Universidade de Lisboa

Faculdade de Ciências
Departamento de Biologia Animal

Microsatellite characterization and marker development from massive sequencing data of the blenny *Salaria pavo*

Sara de Jesus Dias Cardoso

Mestrado em Bioinformática e Biologia Computacional
Especialização em Bioinformática

2011

Universidade de Lisboa

Faculdade de Ciências
Departamento de Biologia Animal

Microsatellite characterization and marker development from massive sequencing data of the blenny *Salaria pavo*

Sara de Jesus Dias Cardoso

Dissertação Orientada pelos Profs. Doutores David Gonçalves (ISPA – Instituto Universitário) e André Falcão (FCUL)

Mestrado em Bioinformática e Biologia Computacional
Especialização em Bioinformática

2011

Agradecimentos

Em primeiro lugar quero agradecer ao meu orientador Dr. David Gonçalves, por me ter integrado no seu projecto de investigação, pelos conhecimentos transmitidos, pelo apoio dado ao longo de todo o trabalho e por me ter dado a conhecer uma espécie singular como são os pavos.

Ao Dr. Rui Oliveira, pela integração no seu grupo de investigação, apoio e bom humor sempre presente em todos os *labmeetings*.

Quero também agradecer ao Dr. André Falcão por ter aceite ser meu orientador e pelo apoio e interesse demonstrado ao longo do meu trabalho.

A todos os elementos do IBBG, que me acompanharam ao longo deste ano, pelo apoio e amizade. À Margarida, pela ajuda inicial que me deu no laboratório e nos primeiros PCRs e à Ana Sofia por nunca me ter deixado ficar sem material para trabalhar. À Silvia pelo companheirismo e boa disposição no tempo que passei na Culatra e no laboratório. À Olinda, Tânia, Leonor, Miguel, Alex e Gonçalo pelos conselhos e ajuda na integração do grupo. À Rute, Ana Isa e Magda pela companhia e pelos bons momentos que proporcionaram no laboratório e nas pausas do café ao longo de todo o ano. À Ana Faustino pelo apoio e ânimo proporcionados e nunca ter levado a mal os meus momentos de mau humor.

Ao Dr. Vitor Almada pelos conselhos dados na análise dos microssatélites e à Dra. Joana Robalo por me ter dado as amostras de pavos das populações mediterrânicas.

Aos meus pais pelo seu apoio, carinho e compreensão, especialmente durante a escrita da dissertação. Ao meu irmão, por todo o apoio que me deu ao longo da vida, pelos conselhos e momentos de descontracção e por me ter deixado "viver" no seu sofá para escrever a dissertação.

Aos meus avós, por terem proporcionado que eu estudasse aquilo que sempre quis.

Resumo

No blenídeo *Salaria pavo* o comportamento reprodutor de fêmeas e machos encontra-se modulado de acordo com a disponibilidade de ninhos no seu habitat. O sistema de acasalamento é promíscuo e os cuidados parentais às posturas são prestados exclusivamente por parte do macho. Nas populações de substrato rochoso onde existe uma grande disponibilidade de ninhos, os machos nidificam em cavidades na rocha e cortejam activamente as fêmeas, enquanto as fêmeas assumem um papel mais passivo. Por outro lado, a população da Ria Formosa apresenta consideráveis modificações ao padrão observado em costas rochosas. Nesta lagoa costeira, os substratos de nidificação são escassos e os únicos ninhos existentes são encontrados em tijolos utilizados na delimitação de áreas de cultivo de bivalves. A escassez de ninhos leva a que nesta população haja uma reversão dos papéis sexuais, com as fêmeas a cortejarem intensamente os machos e a competirem entre si pelo acesso a ninhos, e ao surgimento de tácticas alternativas de reprodução, com o aparecimento de machos parasitas. Estes machos apresentam um tamanho menor daqueles que nidificam e imitam tanto a morfologia como o comportamento de corte das fêmeas de modo a conseguirem aproximar-se do ninho e fertilizar as posturas durante episódios de desova. Estas tácticas alternativas de reprodução são sequenciais e os machos que apresentam uma táctica parasita numa época de reprodução normalmente adquirem um ninho na época seguinte.

De forma a perceber a evolução e manutenção deste tipo de sistemas é necessário estimar o número de ovos que são fertilizados por machos parasitas. Actualmente a forma mais eficaz de fazer testes de paternidades é usando marcadores genéticos, que podem ou não ser complementados com observações de campo. De entre os vários marcadores existentes, o mais utilizado para este tipo de estudos é o microssatélite devido à elevada reproductibilidade dos resultados obtidos.

Os microssatélites são motivos de 1 a 6 nucleótidos repetidos em tandem, altamente polimórficos e facilmente reproduzidos por PCR (*Polymerase Chain Reation*). Até recentemente, o isolamento de novos microssatélites para uma espécie dependia essencialmente da construção de bibliotecas genómicas enriquecidas para determinados microssatélites ou da reutilização de microssatélites de espécies próximas, tendo ambos os processos taxas de insucesso elevadas. Com o rápido desenvolvimento e disponibilização das tecnologias de sequenciação massiva à comunidade científica, uma nova forma de detectar e isolar microssatélites surgiu, a pesquisa *in silico*. De entre as várias plataformas de sequenciação massiva, a mais indicada para espécies que ainda não tenham sequências referência (genoma), é a pirosequenciação, por sequenciar fragmentos suficientemente longos (≈400 pb) que permitem a assemblagem *de novo*.

Neste trabalho, o objectivo principal foi o de isolar e determinar o polimorfismo dos primeiros marcadores genéticos desenvolvidos para esta espécie. Para tal o transcriptoma deste blenídeo, que já se encontra sequenciado por pirosequenciação mas ainda não disponível nas bases de dados públicas, foi utilizado em conjunto com uma ferramenta bioinformática de pesquisa de microssatélites. Cerca de 640,000 sequências foram alinhadas de forma a se obter as 62,038 sequências *consensus* (unigenes com um tamanho médio de 452 pb) que foram utilizadas para fazer uma anotação funcional do transcriptoma e para a pesquisa de microssatélites.

A anotação funcional foi realizada recorrendo a um algoritmo de alinhamento de sequências e procura de similaridades, BLAST (*Basic Local Alignment Search Tool*), mais concretamente o BLASTX, uma vez que se utilizou sequências nucleótidicas para pesquisas de semelhança na base de dados não redundante de sequências proteicas do NCBI. Utilizando um *e-value*$<10^{-5}$, apenas 31% dos unigenes ficaram anotados funcionalmente e 21% tiveram termos do *Gene Ontology* atribuídos. Apesar de a percentagem de unigenes anotados ter sido baixa, a anotação realizada tem significado biológico uma vez que 80% das anotações obtidas foram provenientes de dezoito espécies de peixes.

A pesquisa por microssatélites *in silico* resultou em 4,190 microssatélites identificados em 3,670 unigenes, usando como parâmetros de pesquisa microssatélites perfeitos com um mínimo de 6 repetições para todos os tipos de microssatélites (di-, tri-, tetra-, penta- e hexanucleótidos). Como acontece noutras espécies de peixes, os dinucleótidos são o tipo de microssatélites que se encontram em maior frequência (79%), seguidos dos trinucleótidos (19%). Os restantes tipos de microssatélites encontram-se em menor frequência correspondendo a apenas 6.5% do total de microssatélites.

Qualquer que seja o processo utilizado para a obtenção de microssatélites, será sempre necessário testá-los e aplicá-los num conjunto de amostras de ADN de forma a poder avaliar o seu polimorfismo. Uma vez que neste trabalho interessava isolar microssatélites polimórficos, foram aplicadas duas estratégias de selecção dos microssatélites. A primeira estratégia utilizada baseou-se no conhecimento *a priori* do grau de polimorfismo dos microssatélites na população. Para tal, as sequências individuais que compõem o unigene na região do microssatélite foram manualmente curadas *in silico* de forma a avaliar o seu grau de polimorfismo. No total, 737 microssatélites revelaram ser polimórficos, dos quais 727 eram dinucleótidos e 6 trinucleótidos, com um número máximo de alelos observado *in silico* de 4 alelos. Para além do grau de polimorfismo também se teve em consideração nesta estratégia se o microssatélite estava num unigene anotado e com termos GO atribuídos e se era possível desenvolver um par de primers que o amplificassem. Depois da filtragem dos microssatélites pelos parâmetros mencionados anteriormente obteve-se uma lista de 97 dinucleótidos dos quais 33 foram seleccionados para aplicação (média de 8.5 repetições). A segunda estratégia para a selecção de microssatélites teve

exclusivamente em conta o tamanho da repetição do microssatélite. Estudos anteriores apontam que microssatélites com mais repetições tendem a ser mais polimórficos e, desta forma, foram seleccionados 29 microssatélites para aplicação que continham em média 12 repetições.

Para além dos 63 microssatélites seleccionados, outros 3 microssatélites isolados em *Lipophrys pholis*, pertencente à mesma subfamília do blenídeo *Salaria pavo*, foram também testados numa amostra de ADN de *Salaria pavo* para testar a sua amplificação heteróloga. De modo a testar a viabilidade de aplicação destes microssatélites para estudos de genética de populações, quarenta e um microssatélites, cujos primers amplificaram um só fragmento com o tamanho esperado, tiveram o seu primer *forward* marcado com fluorescência para a genotipagem de 20 indivíduos provenientes da população da ilha da Culatra (Portugal) e 6 indivíduos provenientes das ilhas de Formentera (Espanha) e Borovac (Croácia).

Depois de analisados os resultados obtidos, 28 microssatélites isolados em *Salaria pavo* e 1 microssatélite isolado em *Lipophrys pholis* ficaram validados para futuros estudos. Na população da Culatra todos os microssatélites, à excepção de 5, microssatélites revelaram ser polimórficos. O número de alelos variou entre 2 e 12 alelos e a heterozigosidade observada e esperada variou entre 0.05 a 0.85 e 0.05 a 0.79 respectivamente. O número médio de alelos e a heterozigosidade esperada foi superior nos microssatélites seleccionados usando a segunda estratégia (6.5 e 0.62) comparativamente à primeira estratégia (3.54 e 0.40). Dois microssatélites revelaram estar em desequilíbrio de Hardy-Weinberg e 2 pares de microssatélites em desequilíbrio de linkage. Todos os microssatélites amplificaram nas amostras de ADN das populações de Formentera e Borovac.

Tendo em conta os resultados obtidos, a segunda estratégia revelou ser mais eficiente para a selecção de microssatélites com maior taxa de polimorfismo. Apesar de os microssatélites monomórficos encontrados na população da Culatra terem sido isolados com base na primeira estratégia será necessário aumentar o número de indivíduos genotipados de forma a confirmar-se o grau de polimorfismo observado *in silico*.

Palavras-chave: *Salaria pavo*, Microssatélites, Pirosequenciação, Polimorfismo, Prospecção de dados

Abstract

Next-generation sequencing is providing researchers with a relatively fast and affordable option for developing microsatellite loci for non-model organisms. The number of studies using this approach is fast-growing and a new focus has been given to the development of microsatellites from cDNA due to their potential in targeting candidate genes (type I markers). When the microsatellite polymorphism is of interest, developing microsatellites can become time-consuming due to the numerous primer pairs to be tested for polymorphism by polymerase chain reaction (PCR) in the focal species. Assemblies have a new potential not yet fully explored for microsatellite mining and evaluation, which can help improve the polymorphism rates obtained. Their high sequence coverage enables to access the microsatellite polymorphism *in silico*, if the DNA library sequenced was obtained from a pool of DNA from various individuals of the focal species. Therefore, in this study the transcriptome assembly obtained with pyrosequencing for the blenny *Salaria pavo*, was mined for microsatellites and their polymorphism manually evaluated *in silico*. Two strategies emerged for microsatellite selection and application in a sample of 26 individuals from the islands of Culatra, Formentera and Borovac. Microsatellites were selected based on their *in silico* polymorphism and annotation results (first strategy) or based only on their repetition length (second strategy). From a set of 63 microsatellite loci isolated in *Salaria pavo* sequences, 28 were validated plus one microsatellite from *Lipophrys pholis*. All microsatellites, except 5, revealed to be polymorphic on the 20 individuals genotyped from Culatra Island, the focal population of study.

With the results obtained in this work, the second strategy revealed to be more efficient in yielding polymorphic microsatellites than the first strategy (average number of alleles was 6.5 and 3.54 respectively). Nevertheless, merging these two strategies in future studies may help improving the polymorphism results and at the same time develop type I markers.

Keywords: *Salaria pavo*, Microsatellites, Pyrosequencing, Polymorphism, Data mining

Contents

List of Figures

List of Tables

1. Introduction

This work emerged from the necessity of developing the first genetic markers for the species *Salaria pavo*. Fish species exhibit a high degree of flexibility in their reproductive patterns, and *Salaria pavo* is no exception. This species presents alternative reproductive tactics in habitats where nest availability is low, as it is the case of the focal population of study in Portugal. The main objective here is to use sequences obtained with the new technologies of massive sequencing to isolate new genetic markers for future applications on paternity assessments. Therefore, the markers developed should preferably have high variability to be useful in this type of studies. This matter proves to be highly interesting for bioinformatic analysis, because it is required to work and mine a great quantity of sequence data.

This work was developed in the Integrative Behavioural Biology Group (Unidade de Investigação em Eco-Etologia), whose laboratory is located in ISPA - Instituto Universitário. The primer sequences mentioned in this study haven't yet been published.

1.1 The species *Salaria pavo*

Salaria pavo (Risso, 1810; Teleostei: Blennidae) is a small intertidal fish, usually found in rocky shores of the Mediterranean and adjacent Atlantic areas [1]. This species has a strong sexual dimorphism: males are larger than females and have well-developed secondary sexual characters, namely a conspicuous head crest and a pheromone and antibiotic producing anal gland in the first two rays on the anal fin [2] (Figure 1). In this blenny the mating system is usually promiscuous where the same male may spawn with several females throughout the breeding season and the female may lay her eggs with more than one male. Males build nests (bourgeois males) in holes or crevices in the rock and defend courting territories around the entrance of the nest from which they attract females during the breeding season [1]. Females attach demersal eggs in a monolayer to the cavity walls and males provide uniparental paternal care to the eggs until they hatch [2], which consists of nest defence, nest cleaning and egg fanning. In the peacock blenny populations the reproductive behaviour is modulated according to the nest availability in their habitats. In Mediterranean rocky shore populations nests are available in abundance. Males establish nests in rock crevices or holes, aggressively defend a territory around the nest and actively court females [2]. In contrast, a population at Ria Formosa coastal lagoon (Algarve, Southern Portugal) inhabits an extensive intertidal mudflat area, where the only hard substrates available are agglomerated bricks and tiles used by clam culturists to delimit the frontiers of their fields [3]. Because of the scarcity of nest sites, a strong intra-sexual competition between males is present and only large competitive males are able to acquire and defend a nest. After the breeding season starts, males rarely leave their nests and do not defend

any area around the nest (the breeding territories are restricted to the nest itself), and it is common to observe males nesting in adjacent holes of the same brick [3]. At the peak of the breeding season most nests are filled with eggs and nest space may become a limiting factor for female reproduction [3]. The environmental constrains promote two peculiarities in the mating system of this population: sex-role reversal, where females have the most active role in courtship, and alternative reproductive tactics (ARTs), presence of parasitic males (sneakers) in the population [4]. Although both sexes mate with multiple mates, males are selective in respect to the females they accept as mates. Females, in order to compete for the access to nesting males, approach them with a typical nuptial coloration, consisting of a pattern of light brown and dark vertical bars in the anterior portion of the body and head, that can be turned on/off within seconds, along with elaborated courtship behaviours, involving flickering the pectoral fins and opening-and-closing the mouth in synchrony [4]. Small males are unable to acquire nests and adopt alternative reproductive tactics, acting as sneakers. They approach nest-holder males mimicking the female's morphology and courtship displays in order to come close to the nesting male nest and parasitically fertilize part of the eggs [5,6]. Whenever a parasitic male "steals" some fraction of fertilizations from a nesting male, the nest tender then becomes a foster parent and is said to be cuckolded [7]. The female-mimicry seems to be efficient as nesting males court and attack small sneakers and females with equal frequency [6]. These alternative reproductive tactics are sequential as parasitic males later develop into nesting males after their first breeding season. Since some males do not seem to breed in their first year, these data suggest a condition-dependent tactic for small males that can either reproduce as sneakers or post-pone reproduction to subsequent breeding seasons has nest holders [8].

Figure 1 - *Salaria pavo* main morphotypes.

Sexual dimorphism is evident between the nesting male (top) and the female (middle). Also the intra-sexual polymorphism can be observed between the nesting male and the sneaker male (bottom). Photo by David Gonçalves.

The above morphological descriptions are considered the main morphotypes of this species: nesting male (bourgeois male), female and sneaker male (parasitic male; Figure 1). It is also possible to distinguish at least a third male morphotype, the 'transitional' male, an intermediate stage of the nesting male (sexually immature but phenotypically similar to the nesting male, i.e. sneakers that are already undergoing the change into nesting males) [9].

Estimation of eggs parasitically fertilized is important to understand the evolution and maintenance of alternative reproductive tactics. Indirect evidence for nest parasitism by males can emerge from morphological assessments and behavioural monitoring, but marker-based genetic analyses have made it possible to quantify actual rates and patterns of reproductive cuckoldry in nature. Microsatellite markers, in particular, have proved invaluable for detecting and quantifying reproductive behaviours in fishes, including alternative reproductive tactics and mating systems [10].

1.2 Microsatellites

Microsatellites, also known as Short Tandem Repeats (STRs) or Simple Sequence Repeats (SSRs) are DNA segments usually comprising tandemly repeated motifs of 1 to 6 nucleotides. They belong to a class of highly mutable genomic sequences known as variable number of tandem repeat (VNTR) elements, with rates of mutation higher than the rest of the genome (mutation rate between 10^{-2} and 10^{-6} mutations per locus per generation [11]). They are ubiquitous in eukaryotic and prokaryotic genomes [12,13], present non-randomly[1] in both coding (exonic) and non-coding (intergenic and intronic) regions [14,15], abundant, codominant with Mendelian inheritance and highly reproducible using a simple technique of molecular biology known as Polymerase Chain Reaction (PCR)[2]. Microsatellites are multiallelic (generally have higher heterozygosity), versatile and are more informative than other genetic markers [16]. This makes them powerful genetic markers for genome mapping and for an impressive range of biological questions, from the level of the individual (identity, sex), family (parentage, relatedness), population (genetic structure, epidemiology) and species (phylogenetics, conservation) [17].

[1] All types from (mono- to hexanucleotide repeats) were found to be in excess (compared to random appearance) in noncoding genomic regions across seven eukaryotic clades, suggesting the existence of a strong selection against frame-shift mutations in coding regions resulting from length changes in nontriplet repeats [15].

[2] The purpose of a PCR is to produce millions of copies of a specific DNA sequence. There are three major steps in a PCR, which are repeated for 30 to 40 cycles. In the first step (denaturation), the double strand melts open to single-stranded template DNA (sstDNA), stopping all previous enzymatic reactions (if it isn't the first cycle). In the annealing step, DNA-DNA hydrogen bonds are formed between the primer sequence and the template sequence (when they match). The polymerase binds to the primer-template hybrid and begins DNA synthesis. During extension (the final set of a cycle), DNA polymerase synthesizes a new DNA strand complementary to the DNA template strand by adding dNTPs (deoxynucleoside trisphosphate) that are complementary to the template in 5' to 3' direction.

However, a major drawback of microsatellites is that for most species there is little or no sequence data. Until recently, the advantages of microsatellite markers were partially offset by the difficulty inherent in marker development, as the most commonly used approaches relied on screening of genomic libraries for microsatellite development or testing known microsatellites primers already developed for close species (cross-species microsatellites) [16]. The former demands the construction of a partial genomic library enriched for repetitive motifs, cloning, hybridization to detect positive clones, plasmid isolation and Sanger sequencing followed by primer design and evaluation [18]. Although most of these steps involve relatively straightforward protocols, they can be time consuming due to the frequent need for troubleshooting. Moreover, established protocols for enrichment and cloning can be highly inefficient. For example, the yield of positive clones typically averages only around 2-3% but can fall to as little as 0.03% [18]. The latter may also be unproductive because for most cases the primers available are for non-coding regions, regions where sequences are less conservative [19].

For species with known genome sequences, *in silico* mining of genome databases using bioinformatic tools can be used to identify microsatellites and to design primers targeting these regions [12]. However, genome sequences are available for relatively few eukaryotes and providing these is generally beyond the limited budgets of most research programs. The recent appearance of next-generation sequencing platforms, like Roche's GS-FLX (454 Life Sciences) [20], prove to be an alternative for microsatellite isolation. A single 454 run is capable of generating a large amount of sequence data, with individual expressed sequence tags (ESTs) long enough to capture individual microsatellites along with enough flanking sequence to design PCR primer. Sequence generation on such a scale bypasses the need for enrichment because even a fraction of a 454 run can yield a sufficiently large number of random sequence reads to contain many thousands of microsatellites by chance [21-23].

So, development of microsatellites from genomic libraries is limited to those motifs for which the initial hybridization or enrichment was performed and in most cases the PCR primers used to amplify the microsatellites are species-specific, which implies that markers developed in one taxon cannot be easily and readily applied to others. Normally these microsatellites are considered Type II markers, because they are associated with genomic regions that have not been annotated to known genes [24]. On the other hand, expressed sequencing tag (EST) derived microsatellites (EST-SSRs) are less polymorphic due to functional restraints [25], compared to those derived from non-coding genomic sequences, but have several intrinsic properties making their identification desirable. Because EST-SSRs are exonic, their flanking regions are expected to be more conserved across closely related species and with a lower rate of null allele appearance [19,26,27]. These microsatellites represent a potential source of Type I

markers, which are linked to genes (of known function) [24], making them more useful for comparative genetic mapping, linkage and quantitative trait loci (QTL) association studies.

1.3 Pyrosequencing and Functional Annotation

Next-Generation DNA sequencing refers to various technologies that implement cyclic-array sequencing (for a review see [28-30]). The concept of cyclic-array sequencing can be summarized as the sequencing of a dense array of DNA features by iterative cycles of enzymatic manipulation and imaging-based data collection [31].

Sequencing entire genomes of non-model organisms is still out of reach for most researchers but sequencing smaller subsets of the genome, like transcriptomes, provides an attractive alternative. Transcriptomes correspond to the transcribed DNA of an organism and therefore represent functional genomic data. De novo assembly and annotation is easier for transcribed genes than for complete genomes because new sequences can be compared to conserved protein sequences and transcribed genes contain fewer repetitive elements [32]. Creating a reference transcriptome can be an invaluable tool for deciphering the genetic architecture of adaptive traits in species for which complete genome sequence is not available and for Type I genetic marker development.

The 454 system, also known as pyrosequencing, was the first next-generation sequencing platform available as a commercial product [20]. Initially, pyrosequencing was restricted to model organisms because of the short reads (100-200 bp) produced that made *de novo* assembly difficult without a reference genome. Nowadays, it offers the possibility of long sequencing reads (longer than the other systems [29]), with more accurate base calling and deeper sequencing coverage, important for *de novo* transcriptome assembly of the ESTs, meaning that transcribed genes of non-model organisms can be characterized without a pre-existing sequence reference (genome) [33-35].

In the pyrosequencing process, one nucleotide at a time is washed over several copies of the sequence to be determined, causing polymerases to incorporate the nucleotide if it is complementary to the template strand. The incorporation stops if the longest possible stretch of complementary nucleotides has been synthesized by the polymerase. In the process of incorporation, one pyrophosphate per nucleotide is released and converted to ATP by an ATP sulfurylase (Figure 2). The ATP drives the light reaction of luciferases present and the emitted light signal is measured. The amount of light produced is proportional to the number of nucleotides incorporated (up to the point of detector saturation). After capturing the light intensity, the remaining unincorporated nucleotides are washed away and the next nucleotide is provided.

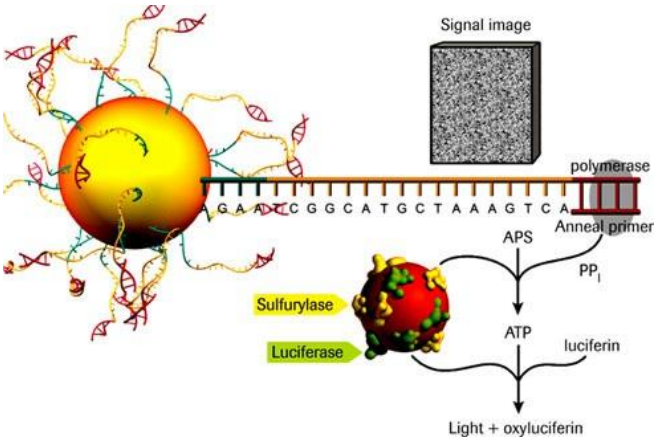


Figure 2 - Pyrosequencing chemistry.

Single stranded DNA template linked to a capture bead (yellow) is exposed to only one nucleotide during each round of sequencing. As a nucleotide (thymine in this example) is incorporated through the action of DNA polymerase (brown), inorganic pyrophosphate (PPi) is released which reacts with adenosine 5'-phosphosulfate (APS) and sulfurylase to generate ATP. ATP is then used as a substrate for luciferase to generate light, which can be detected and quantified. Adapted from 454 Life Sciences - The Technology (http://my454.com/products/technology.asp).

For the library construction, DNA samples are nebulized into small fragments and short specific adaptors (A and B) are ligated onto 5' and 3' end extremities of each fragment (Figure 3a). The library fragments are mixed with a population of agarose beads whose surfaces carry oligonucleotides complementary to the 454-specific adapter sequences on the fragment library, so each bead is associated with a single fragment (Figure 3b). Each of these fragment-bead complexes is isolated into individual oil-water micelles, that also contain PCR reactants, and thermal cycling (emulsion PCR) of the micelles occurs producing approximately one million copies of each DNA fragment on the surface of each bead.

These amplified single-stranded molecules are then sequenced en masse (Figure 3c). First the beads are arrayed into a picotiter plate (PTP; a fused silica capillary structure) that holds a single bead in each of several hundred thousand single wells, which provides a fixed location at which each sequencing reaction can be monitored. Enzyme containing beads that catalyze the downstream pyrosequencing reaction steps are then added to the PTP and the mixture is centrifuged to surround the agarose beads. On instrument, the PTP acts as a flow cell into which each pure nucleotide solution is introduced in a stepwise mode (cycles), with an imaging step after each nucleotide incorporation cycle. The PTP is seated opposite a CCD (charged-coupled device) camera that records the light emitted at each bead as corresponding to the array coordinates of specific wells. The first four nucleotides (TCGA) on the adapter fragment adjacent to the sequencing primer added in library construction correspond to the sequential flow of nucleotides into the flow cell. This strategy allows the 454 base-calling software to calibrate the light emitted by a single nucleotide incorporation. The sequencing is

'asynchronous' in that some features may get ahead or behind other features depending on their sequence relative to the order of base addition.

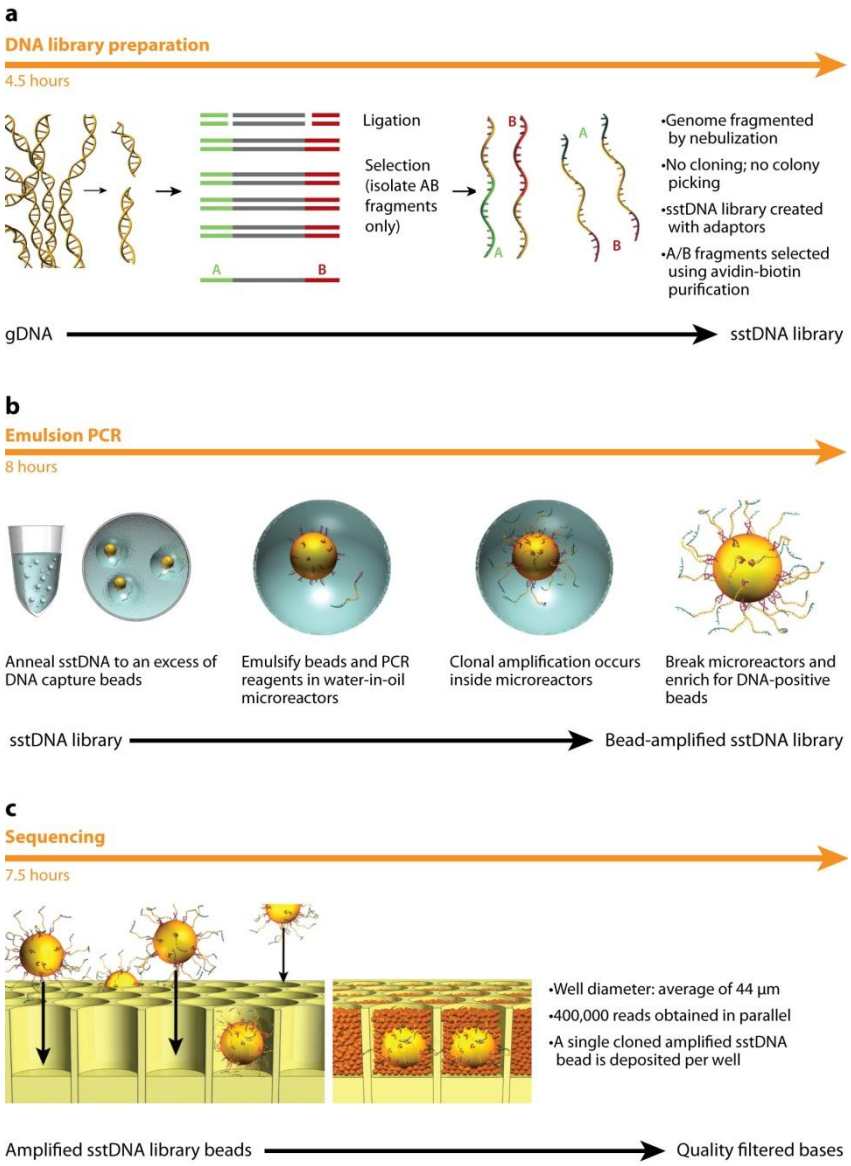


Figure 3 - Workflow of Roche/454 sequencing platform.
DNA is fragmented using nebulization and to its extremities adaptors are ligated (**a**); a mixture of DNA fragments with agarose beads containing complementary oligonucleotides to the adapters are mixed in an approximately 1:1 ratio and encapsulated by vigorous vortexing into aqueous micelles that contain PCR reactants surrounded by oil, and pipetted into a 96-well microtiter plate for PCR amplification (**b**); the resulting beads are decorated with approximately 1 million copies of the original single-stranded template DNA (sstDNA), which provides sufficient signal strength during the pyrosequencing reaction that follows to detect and record nucleotide incorporation events. Adapted from Mardis [28].

A major limitation of the 454 technology relates to homopolymers (consecutive instances of the same nucleotide). The calibrated base calling cannot properly interpret long stretches (>6)

of the same nucleotide, so these areas are prone to base insertion and deletion errors. In contrast, because each incorporation step is nucleotide specific, substitution errors are rarely encountered. Most of these problems can be resolved by higher read coverage.

The current 454/Roche GS FLX Titanium platform makes it possible to sequence about 1.5 million such beads in a single experiment and to determine sequences of length between 300 and 500 bp [29]. The length of the reads is determined by the number of flow cycles (the number of times all four nucleotides are washed over the plate) as well as by the base composition and the order of the bases in the sequence to be determined. Currently, 454/Roche limits this number to 200 flow cycles, resulting in an expected average read length of about 400 bp. This is largely due to limitations imposed by the efficiency of polymerases and luciferases, which drop over the sequencing run, resulting in decreased base qualities.

All the sequence data gathered must be functionally annotated after quality control and assembly of the reads obtained, in order to assign them a function by homology searches to known sequences. BLAST (Basic Local Alignment Search Tool) is a NCBI's (National Center for Biotechnology Information) sequence similarity search tool designed to support analysis of nucleotide and protein databases [36]. BLAST is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA.

Given the choice of aligning a DNA sequence or the sequence of the protein it encodes, it is often more informative to compare protein sequences. One of the reasons for this is that many changes in a DNA sequence (particularly at the third position of a codon) do not change the amino acid that is specified, allowing protein sequence comparisons with sequences from organisms that last shared a common ancestor over 1 billion years ago [37].

For the functional annotation, the sequences of interest must be in FASTA format, which consists of a short description line preceded by a "greater than" (">") symbol and the sequence itself in the next lines. Also it is necessary to select a database to which the sequences are to be queried. At NCBI there are several protein and nucleotide sequence databases that can be searched with the right BLAST program. One of these databases is the non redundant protein sequence database that can be searched using BLASTX [38]. This database is a non redundant collection of protein sequences from the following protein databases: non-redundant GenBank coding sequences, PDB (Protein Data Bank), SwissProt, PIR (Protein Information Resource) and PRF (Protein Resource Foundation). BLASTX translates a DNA sequence into six protein sequences using all six possible reading frames, and then compares each one of these proteins to the proteins sequences. To enrich the information gathered for the sequences of interest, they can be also annotated with a structured controlled vocabulary. The Gene Ontology (GO) project [39] is a major bioinformatic initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. It is composed of three ontologies that

describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner.

The similarity searches can be done at the NCBI website, locally using blast+ or using a platform specialized in annotation of sequences. The last resource can become more useful for data annotation, because these platforms normally have inbuilt workflows or strategies specifically for this. One example of such platform is Blast2GO [40], a web platform with the main purpose of enabling Gene Ontology (GO) annotation based on similarity searches with statistical analysis in non-model species.

1.4 Objectives

In this work, the main objective was to identify microsatellite loci for *Salaria pavo*, a species that currently hasn't any markers available.

The transcriptome of this species has been recently sequenced using Roche's 454 platform as part of an ongoing FCT project coordinated by the research team (PTDC/MAR/69749/2006), and as such the sequences obtained aren't yet available in the public biological databases.

To achieve the primary goal, the assembled ESTs were functionally annotated and mined for microsatellite content. To facilitate gene identification, the assembled sequences were annotated using the structured vocabulary provided by the Gene Ontology Consortium, based on the results obtained with BLAST searches.

Independently of the approach used for marker development, it will always be necessary to design primers, optimize each primer pair for PCR and then test these for polymorphism in a sample of individuals. Following this line of thought, two strategies emerged for microsatellite selection and application in order to get improvements in efficiency in obtaining polymorphic microsatellites:

- The first strategy used a pre-screening of microsatellites for polymorphism evaluation *in silico* and type I marker development.
- The second strategy intended to test the relationship between the microsatellite repeat length and its degree of polymorphism.

2. Material and Methods

The start point of the work developed for this thesis begins in method 2.1.3. The previous methods described were already completed, but it is important to mention them in order to get a perspective of how the sequence data was obtained and its nature.

In a general perspective of the work done, the assembly of the sequences obtained by pyrosequencing was first annotated and then characterized for its microsatellite content. A set of selected microsatellites was then applied in DNA samples from three different populations of *Salaria pavo*. Finally, the results obtained were analysed according to different population genetic analyses.

2.1 Transcriptome sequencing and analysis

2.1.1 Sampling and cDNA library construction

Peacock blenny tissue samples were taken from 13 individuals (3 males, 3 females, 3 sneakers and 4 transitional males) sampled at Culatra Island, Ria Formosa (36°59′N, 7°51′W, Algarve; for a detailed description of the area of use see [3]). Total RNA was separately isolated with TRI Reagent® (Sigma-Aldrich) following standard procedures from 12 tissues including skin, muscle, bone, brain, olfactory epithelium, eyes, heart, kidneys, spleen, intestine, gonads and anal gland. The purity and quantity of the resulting RNA was determined using ND-1000 spectrophotometer (Nanodrop Technologies, Wilmington, DE) and gel electrophoresis.

Equal masses of total RNA from these tissues were pooled and used to construct one normalized cDNA library at the Max Planck Institute of Molecular Genetics (Berlin, Germany). The SMART (Switching Mechanism At 5' end of RNA Template) PCR cDNA kit (Clontech, Palo Alto, CA, USA) [41] was used to construct the cDNA library and Poly(A+) mRNA was isolated from total RNA using the PolyATtract® mRNA Isolation System (Promega, Madison, WI, USA). This library was later normalised using the duplex-specific nuclease (DSN) method [42] in order to optimize the random sequencing from cDNA library, by equilibrating the abundant and rare transcripts.

2.1.2 Pyrosequencing and cluster assembly

Sequencing and cluster assembly was carried out at Max Planck Institute of Molecular Genetics (Berlin, Germany). Sequencing was performed using GS FLX Titanium series reagents and using one single region on a Genome Sequencer FLX instrument (Roche/Life Sciences 454). Bases were called with 454 software by processing the pyroluminescence intensity for each bead-containing well in each nucleotide incorporation.

The resulting ESTs were quality-trimmed (≥q20) to remove adapter sequences, vector clipped using Phred [43,44] and LUCY [45] with standard parameters, and all sequences shorter than 100 bp were discarded. Repeats were masked by RepeatMasker [46] before the ESTs were clustered and assembled *de novo*, applying the "accurate" mode with default parameters, as defined in MIRA3 assembler [47]. Sequences after assembly comprised contigs ('c'), repetitive contigs ('lrc'), singletons ('s') and ''debris" reads: consensus sequences originating from clusters of at least two reads representing the same transcripts were termed contigs or repetitive contigs if the building of the new contig was done in a repetitive region; unique sequences that clustered with other reads, but that were eventually excluded from the final assembly were classified as singletons; and ESTs that may be of high-quality, but display no significant relationships to any other reads during the assembling process were classified as ''Debris" reads.

2.1.3 Homology searches and assembly annotation

The BLASTX algorithm [38] was used to query for sequence similarity on all *S. pavo* transcripts (contigs, repetitive contigs and singletons) against the NCBI non redundant protein sequence database (nr, release of May 2010, including all non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF) using Blast2GO suite [40]. Homology searches were performed remotely on the NCBI server through QBLAST. Alignments with an E-value $<1e^{-5}$ and a minimum coverage between the hit and the query sequence of 33% up to 10 hits per contig were taken into account. Blast2GO was also used for functional annotation of the transcripts applying the function for the mapping of Gene Ontology (GO) terms [39] to transcripts with BLAST hits obtained from BLASTX searches against nr database. Using default parameters, only ontologies obtained from hits with E-value $< 1e^{-6}$, annotation cut-off > 55, and a GO weight > 5 were used for annotation. The transcripts were also mapped to metabolic pathways in accordance with the KEGG (Kyoto Encyclopedia of Genes and Genomes) [48] and Enzyme Commission (EC) numbers [49] were obtained and used to putatively map unique sequences to specific biochemical pathways.

2.2 Microsatellite mining and application

2.2.1 Microsatellite mining

The identification and localization of perfect microsatellites in the assembled contigs was accomplished using MSATCOMMANDER version 0.8.2 [50]. This program is designed to take as input DNA sequence data in FASTA-formatted files and search for microsatellite motifs and design primers. The parameters were set for detection of di-, tri-, tetra-, penta-, and hexanucleotide motifs with a minimum of six repeats, and the option "Design Primers" was also

chosen, keeping the primer criteria as described in Faircloth [50]. Tab-delimited files were generated from the searches using MSATCOMMANDER, and converted to spread-sheet files for use in Microsoft Office Excel 2007 (Microsoft Corporation, 2007) for subsequent data manipulation as described in Santana *et al.* [51].

Contigs harbouring microsatellites were manually curated with the aid of Tablet version 1.11 [52], an assembly graphical viewer, for existence of polymorphism shown by the reads within the contig. The main display window of this program allows a single contig to be viewed at a time and is navigable by means of scrolling and zooming functions (Figure 4). Within this window, each of the reads is shown aligned against the consensus sequence, with individual bases coloured according to nucleotide type and each read occupying a separate row under the 'stacked format' option. Pad characters, introduced by Mira to fill any gaps in the assembly that arise for example where indels (insertion or deletion sites) or length polymorphisms are present among the reads, are represented by star symbols. With the aid of this program, for each microsatellite information was kept of how many reads covered completely the microsatellite region (read coverage), if the microsatellite was at the 5'end or 3'end of the contig (incomplete microsatellite) and the number of length variants for the microsatellite shown by the reads (alleles, Figure 4). Microsatellites were also functionally annotated with accordance of its contig by using the assembly annotation results previously obtained.

Two different strategies to select microsatellites for application were pursued. The first strategy required the fulfilment of four conditions for a microsatellite to be considered for genotyping: 1) display of polymorphism by the reads forming the contig cluster; 2) presence of BLAST results; 3) GO terms assigned, and finally 4) existence of at least a pair of primers generated by PRIMER3 [53]. In the second strategy, the microsatellites were only selected based on the length of the repetition and the existence of a pair of primers, not limiting the remaining parameters described above. For both strategies, the quality of the flanking regions was evaluated and only one microsatellite per pair of primers was considered, in other words, one pair of primers only amplified one microsatellite.

In order to eliminate any possible redundancy, that may exist in the assembled database, and unspecific amplifications, which could lead to an overestimation of available markers for testing, selected primers were tested against the database using PrimerValidator [54]. This program allows at most one mispaired base between primer and sequence, switching the base to an X in the primer when printing the results.

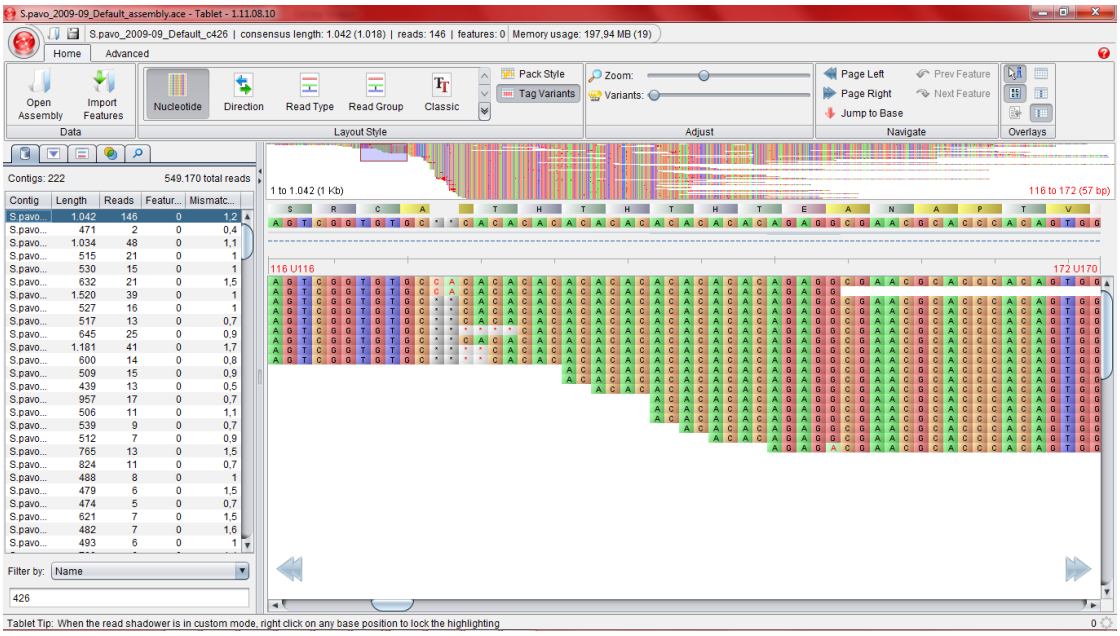


Figure 4 - Screenshot of a polymorphic dinucleotide repeat (CA) microsatellite locus visualised *in silico* using the program Tablet.

The upper 'overview window' shows a summary of all of the reads comprising the consensus sequence, while the main 'display window' shows the microsatellite and its immediate flanking regions visualised under a higher zoom. Within this window, 454 reads are shown aligned against the consensus sequence, with each read occupying a separate row. Individual bases are coloured according to nucleotide type and pad characters, introduced by Mira to fill any gaps in the assembly, are represented by star symbols against a light grey background. Four distinct motif length variants can be seen, comprising twelve, eleven, ten and nine repeat units.

2.2.2 Microsatellite PCR amplification and polymorphism screening

A set of 63 microsatellites developed from *Salaria pavo* contigs and other three microsatellites previously developed for *Lipophrys pholis* [55], a species belonging to the same subfamily of *Salaria pavo*, were selected for amplification test using one peacock blenny DNA sample. PCR amplifications were set up in 50µl volume composed of ~100ng DNA, 0.25 pmol of each primer (MWG), 1.5 mM $MgCl_2$ (for exceptions see Table 6), 120µM of each dNTP[3], 5XGreen GoTaq® Flexi Buffer 1X, and 1.5u *Taq* DNA polymerase (Promega). PCRs were performed in a thermal cycler (Stratagene RoboCycler® Gradient 96) programmed as: 3 min at 94ºC for initial denaturation, followed by 35 cycles of 94ºC for 1 min, primer specific annealing temperature for 1 min, 72ºC for 45 s, and a final extension at 72ºC for 7 min. The success of the PCRs was determined by running 8µL of each PCR product and co-running 5µl of a mixture of DNA loading dye with a 50 bp DNA ladder (GeneRuler™ 50 bp DNA Ladder – 0.5µg/µl, Fermentas) on a 1X Tris-acetate-EDTA buffer and 2% agarose gel stained with GelRed 3X, visualized under UV light and photographically documented.

[3] Deoxynucleoside trisphosphate

For peacock blenny's loci that seemed to amplify well in agarose gels, the respective forward primers were re-ordered labelled 5'-end fluorescently with 6-FAM[4] or with HEX[5] dyes (Table 6). A total of 26 adult peacock blenny individuals sampled from Culatra Island, Formentera Island (Spain, 38° 41'N, 1° 27'E) and Borovac Island (Croatia, 43° 9'N, 16°24'E) (Figure 5) were employed for polymorphism assessment. DNA was extracted from the dorsal fin using Extract-N-Amp™ Tissue PCR Kit (Sigma-Aldrich). Microsatellite amplification reactions were performed in 25µl volume containing ~100ng DNA, 0.25 pmol of each primer (MWG), 1.5 mM $MgCl_2$ (for exceptions see Table 6), 60µM of each dNTP, 5X Green GoTaq® Flexi Buffer 1X, and 0.75u *Taq* DNA polymerase (Promega). PCR thermal programme was run as previously described (for annealing temperatures see Table 4).

Figure 5 - Geographic location of the populations from which samples were used for genotyping. Black circles indicate sample sites of *S. pavo*: Cl_PT – Culatra Island, Portugal; Fm_SP – Formentera Island, Spain; Bv_CR – Borovac Island, Croatia.

At the end of the PCR, all DNA fragments were fluorescently marked because they were amplified using primers labelled by one of the two dyes, which allowed, by fluorescence detection with capillary electrophoresis obtaining the respective fragment size for each DNA sample used in each microsatellite. DNA fragments were separated on a commercial ABI 3730XL DNA analyzer and sized by co-running a GeneScan HD400 (Applied Biosystems) size standard. PCR products were loaded into the capillary array by a short period of electrophoresis called *electrokinetic injection* and separated by size as they travel through the polymer-filled capillary array (electrophoresis). As they reach the detection window, the laser beam excites the dye molecules and causes them to fluoresce. Software converts the banding pattern into a plot with peaks corresponding to the width and intensity (height) of each band. The position of the

[4] 6-carboxyfluorescein

[5] Hexachloro-fluoresceine, a phosphoramidite from Applied Biosystem.

peak along the x-axis corresponds to the size of the DNA product in the band measured in base pairs (bp). The height/intensity corresponds to the concentration of the DNA product, which is a consequence of the efficiency of the amplification process in PCR. One colour is used for a size standard to calibrate the band positions with the size of the DNA product (using HD400 size standard the colour is red).

DNA fragments were scored manually with the aid of GeneMarker® version 1.95 (SoftGenetics, State College, PA, USA). *Salaria pavo* is a diploid fish, so it was expected to see one peak, corresponding to one allele (same number of repeats on both homologous chromosomes) or two peaks, corresponding to two alleles (different number of repeats for homologous chromosomes). To standardize the scoring of fragments sizes, rules were created to analyse the peak intensities based in the fact that two alleles from the same locus aren't always equally amplified during the PCR. This phenomenon, known as 'allelic dropout', can occur due to random preferential amplification of one allele during the PCR, leading to the misidentification of heterozygotes as homozygotes due to reduced peak intensity of the poorly amplified allele [56], or it may also be caused by variation of the flanking region used by a PCR primer so that the primer does not bind properly as in the case of null alleles [57]. In accordance with this, an individual was considered homozygotic for a particular microsatellite locus when in the target range of possible fragment sizes existed only one peak or when in the presence of more than one peak they were smaller than half of the height of the highest peak. An individual was considered heterozytic for a particular microsatellite locus when there were two peaks with equal heights or when the smallest of the two peaks had at least half the height of the highest one (Figure 6). When necessary, the alleles were confirmed by commercial sequencing.

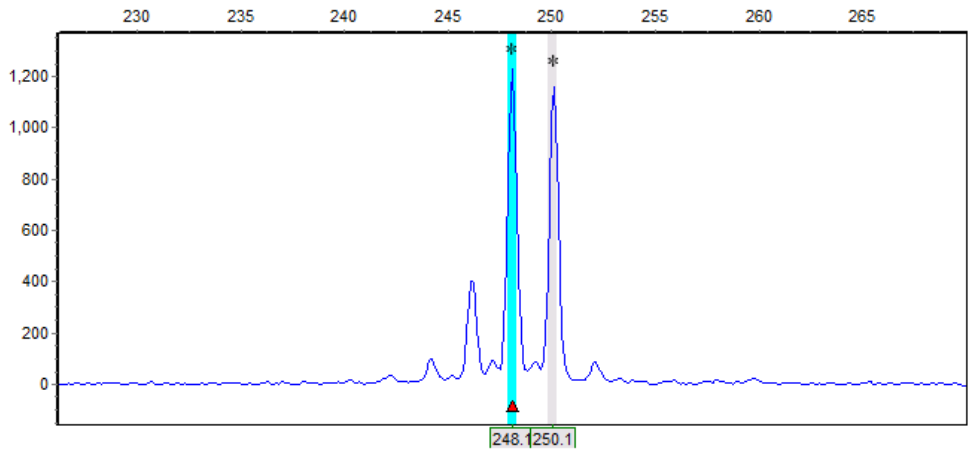


Figure 6 - Screenshot of an output of fragment sizing from GeneMarker program.
For this microsatellite locus the individual being analysed is considered heterozygotic. The two peaks are identified with asterisks and correspond to alleles of fragment size 248 and 250 (dinucleotide), which are almost equal in height. X-axis corresponds to the size of the DNA product in base pairs (bp) and the Y-axis correspond to the intensity (concentration) of the DNA product.

For each working loci the type of repetition and its length was confirmed by commercial sequencing. The fragment length obtained for a DNA sample by sequencing for each microsatellite locus was compared to the respective value obtained by capillary electrophoresis using the same DNA sample. This was done due to the fact that dyes have different motilities during electrophoresis leading to fragment sizes not necessarily corresponding to the actual length determined by direct sequencing [57-59]. If the fragment sizes were different between the two techniques, the values obtained by capillary electrophoresis were corrected to the values obtained by sequencing.

2.2.3 Microsatellite loci evaluation

After the definite set of working microsatellites have been established, several general population genetic measures were inferred to determine the usefulness of the microsatellites developed.

First the gathered data was tested in relation to its conformity to Hardy-Weinberg equilibrium (HWE) and genotypic linkage disequilibrium (LD). The Hardy-Weinberg principle is a model that relates allele frequencies to genotype frequencies and presumes that these don't change from one generation to the next (are in equilibrium), assuming random mating, no mutation, no genetic drift and no migration. Linkage disequilibrium tests the non-random association of alleles at two or more loci in order to avoid pseudo-replication of the analysis in which loci are assumed to be independent samples. Tests for HWE and LD were performed using the GENEPOP version 4.0.11 [60] with the default setting (10,000 dememorization steps, 100 batches, and 5,000 iterations per batch) using complete enumeration method [61] (whenever possible) and Markov chain method [62].

Genetic diversity estimates were obtained by calculating the observed (Ho) and expected (He) heterozygosities. The Ho is the proportion of heterozygous individuals for each particular locus that can be found in the surveyed population. The He is the estimated fraction of all individuals that would be heterozygous for any randomly chosen locus. The He differs from the Ho because it is a prediction based on the known allele frequency from a sample of individuals. These measures were also calculated using GENEPOP.

The test for the presence of null alleles[6], stuttering[7] and large allele drop[8] out was conducted using MICRO-CHECKER version 2.2.3 [63]. This application uses a Monte Carlo

6 Null alleles are non-amplified alleles that, when segregating with another allele, result in an apparent homozygote. For microsatellites, such null alleles can arise when mutations occur in the flanking regions, preventing one or both of the primers from binding. The indication for the presence of null alleles is the general excess of homozygotes and the heterozygote deficit.

7 Stutter is due to mistakes that occur during PCR producing additional products that differ from the original template by multiples of the repeat unit length.

simulation (bootstrap) method to generate expect homozygote and heterozygote allele size difference frequencies.

The program Populations version 1.2.32 [64] calculates various genetic distances for individuals and populations based on the alleles found. This program was used to create distance matrices for individuals with the genetic distance D_{AS} [65] which calculates shared allele distances, and for populations with Cavalli-Sforza and Edwards chord distance D_C [66]. Allele sharing (D_{AS}) is a genetic distance measure based on allele frequencies only, and can be defined as the probability that two alleles randomly chosen from both populations are not identical [67]. Cavalli-Sforza and Edwards chord distance (D_C) makes no biological assumption and the magnitude of this distance is not proportional to evolutionary time, but its use generally leads to a higher probability of depicting the correct tree among closely related populations [68]. All distances were used in conjunction with the "Neighbour Joining" algorithm [69] and all produced trees were manipulated with MEGA 4.0 software [70].

[8] Large allele dropout is the failure to amplify the larger of two alleles in a heterozygote due to the preferential amplification of the smaller allele.

3. Results

3.1 Transcriptome sequencing and analysis

3.1.1 Transcriptome Assembly

After discarding the poor quality sequences, a total of 642,012 reads with an average length of 321 bp were *de novo* assembled into contigs (Table 1). Of these reads, 92,842 high quality reads were not assembled ("Debris" reads) and were excluded from further analysis. The assembly produced 62,038 sequences, with an average length (mean ± standard deviation) of 451.7 ± 171.8 bp (range 100-3,153 bp) and a mode of 466 bp. The longest 10% sequences ranged from 635bp to 3,153 bp (n = 6,204). Of these assembled sequences, 96.2% were contigs (n = 59,683), 3.5% were repetitive contigs (n = 2,155) and 0.3% were singletons (n = 200). These sequences putatively correspond to different transcripts and were designated unigenes. The average number of reads per unigene (without singletons) was 8.9 ± 42.8 (range from 2 to 2,646 number of reads) and the mode was of 2 reads per unigene. Unigenes in the top 10% of number of reads ranged from 12 to 2,646 reads (n = 6,184).

Table 1 - Summary statistics of 454-pyrosequencing assembly.
Type 'c' - contigs, type 'lrc' - repetitive contigs and type's' - singletons

Input 454 reads	642,012
Average read length (bp)	321
Total amount	206.17 Mb
Assembly Results	
Assembled reads	549,170
Debris Reads	92,842
All contigs	62,038
Type 'c'	59,683
Type 'lrc'	2,155
Type 's'	200

3.1.2 Functional Annotation and Gene Ontology Analyses

The resulting unigenes (62,038 contigs) were used to search the NCBI nr database, the most comprehensive and non-redundant collection of proteins, using BLASTX. Nearly 31% of these unigenes (19,008 contigs) had hits in the BLAST searches with a cut-off e-value of less than 10^{-5} (for a detailed view of the BLASTX results, see Appendix A). Totally, 165,615 hits were identified, with an average of 8.82 ± 2.75 hits per unigene and a mode of 10 hits. Table 2

shows the distribution of the significant BLASTX hits of the unigenes per organism. Roughly, 15,201 unigenes had its top BLAST hit from one of 18 fish species, corresponding to 80% of the BLAST results.

Table 2 - Distribution of significant homologous matches ($e<10^{-5}$) of the unigenes per organism. All fish species are highlighted with gray colour (closest species to *Salaria pavo* with dark gray).

Organism	BLAST Top-Hits
Tetraodon nigroviridis	6,581
Danio rerio	3,779
Salmo salar	1,375
Anoplopoma fimbria	887
Osmerus mordax	367
Takifugu rubripes	341
Oryzias latipes	333
Epinephelus coioides	313
Xenopus (Silurana)	254
Oncorhynchus mykiss	187
Esox lucius	163
Paralichthys olivaceus	162
Gallus gallus	142
Ictalurus punctatus	137
Taeniopygia guttata	134
Siniperca chuatsi	120
Solea senegalensis	107
Mus musculus	104
Oreochromis mossambicus	100
Monodelphis domestica	96
Sparus aurata	91
Xenopus laevis	91
Ailuropoda melanoleuca	86
Homo sapiens	83
Oreochromis niloticus	83
Ornithorhynchus anatinus	79
Sus scrofa	77
Pagrus major	75
Unknown	71
Others	2590

BLASTX top-hit species distribution of gene annotations showed highest homology to *Tetraodon nigroviridis* sequences, followed by *Danio rerio* and *Salmo salar* (Table 2),

accounting for 62% of the top BLAST results. The species phylogenetically closer to *Salaria pavo* present in Table 2 (n = 6) belonged to the same order (Perciformes) and only accounted for 4% of the top BLAST results. The other fish species belonged to the same superorder of *Salaria*, as it is the case of *Tetraodon nigroviridis*, or to other divisions inside Teleostei group. Ten species listed belonged to other classes (Mammalia (n = 6), Aves (n = 2) and Amphibia (n = 2)) and accounted for 6% of the top BLAST results. BLASTX result accessions were used to retrieve gene names and GO terms.

Gene Ontology annotation was employed to interpret the function of the unigenes. From the total number of unigenes with BLAST results 15,949 unigenes (83.9%) had GO terms attributed. The average number of GO terms per unigene was 5.40 ± 5.41 terms (range 1-107 terms) with a mode of 3 GO terms. Of these, 13,091 unigenes (82%) were annotated using Blast2GO default parameters, resulting in 70,598 GO terms annotated out of 80,588. The unigene with more GO terms attributed (107 terms) was classified as Catenin beta-1 protein, a subunit of the cadherin protein complex and an integral component in the Wnt signalling pathway.

Combined graphs of GO terms were constructed based on at least 10 unigenes per node in the Directed Acyclic Graph (DAG)[9], meaning that these graphs will contain only those functional terms (GO terms) which are covered by at least 10% of the *S. pavo* unigenes annotated. Within the three major divisions of Gene Ontology, 'Biological process' constituted the largest division of GO assignment of the unigenes (33,153 counts), followed by 'Cellular Component' (19,822 counts) and 'Molecular Function' (18,120 counts). Level 2 GO assignments within the three major divisions are summarized in Figure 7. The first, 'Biological Process', refers to the 'biological objective to which the gene or gene product contributes' [39]. Among the 'Biological Process' GO terms, 20 categories were identified and the two most abundant were 'cellular process' (26% - 8,558 unigenes) and 'metabolic process' (20% - 6,478 unigenes). The second division, 'Molecular Function', refers to 'some biochemical activity that is performed by the gene, without a temporal or spatial context' [39]. In this division, 12 categories were identified and the vast majority of GO terms were involved into binding (47% - 8,446 unigenes) and catalytic activities (28% - 5,044 unigenes). The third division, 'Cellular component', describes 'the sub-cellular location where a gene product is active' [39]. Under this division, 6 categories were identified and 75% of all GO terms corresponded to cell parts and organelle (9,184 and 5,809 unigenes respectively).

[9] The structure of GO can be described in terms of a graph, where each GO term is a node, and the relationships between the terms are arcs between the nodes (network). These relationships are *directed* (*a* is a *b*, but *b* is not an *a*) and the graph is *acyclic*, meaning that cycles are not allowed in the graph. The terms get more specialized going down the graph, with the most general terms (the root nodes: cellular component, biological process and molecular function) at the top of the graph.

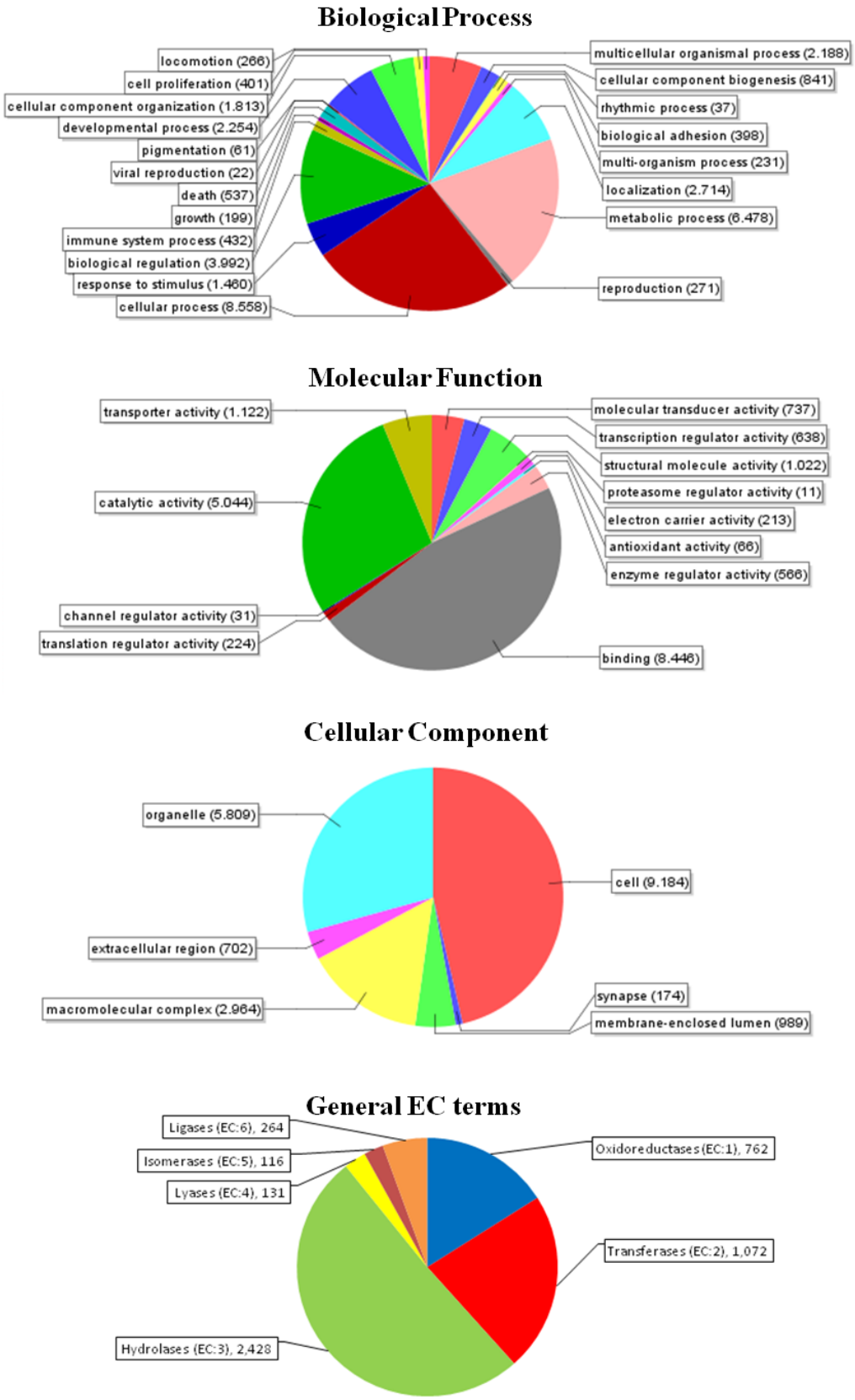


Figure 7 - Gene Ontology (GO) assignment (2nd level GO terms) and Enzyme Classifications (EC) for the *Salaria pavo* transcriptome.

Note that one sequence can be associated with more than one GO term and with more than one EC number.

Within the 4,773 predicted EC numbers (Figure 7), 'Hydrolases' (51% - 2,428 unigenes) was the dominant predicted enzyme function followed by 'Transferases' (22.5% - 1,072). A total of 127 metabolic pathways were obtained from KEGG pathway database (data not shown). The highest number of KEGG mappings was extracted from 'Metabolic Pathways' and 'Biosynthesis of secondary metabolites' with 1213 and 392 sequences respectively.

For a summary of the functional annotation results, see Appendix B.

3.2 Microsatellite mining and application

3.2.1 Microsatellites types and distribution

A complete search in the assembly for 5 types of microsatellites with a minimum repeat length of 6 repeats resulted in the identification of 4,190 microsatellite loci in 3670 unique unigenes, representing 5.9% of the transcriptome (Table 3). With this value, unigene-microsatellite (EST-SSR) frequency obtained for peacock blenny was 6.75%. Of the 4,190 microsatellites, 520 microsatellites were present in unigenes that already had one type of microsatellite (Figure 8).

Table 3 - Summary of the results obtained with *in silico* mining for microsatellites in *S. pavo* unigenes. Type 'c' - contigs, type 'lrc' - repetitive contigs and type's' - singletons

Number of unigenes analysed	62,038
Number of SSRs identified	4,190
In unigene type 'c'	5,4%
In unigene type 'lrc'	44,5%
In unigene type 's'	2,5%
Number of unique unigenes containing SSRs	3,670

The dinucleotide repeats were the most abundant type of microsatellites within peacock blenny unigenes. They accounted for 79.0% of all microsatellite containing unigenes, followed by 14.5% for trinucleotides, 4.4% for tetranucleotides, 1.4% for pentanucleotides and 0.7% for hexanucleotide repeats (Figure 8).

Of the dinucleotide repeats, AC/GT was the most abundant motif, accounting for 78.6% of all dinucleotide repeats found in the unigenes (Table 4). AG/CT was the second most abundant dinucleotide motif type, accounting for 20.1% of all dinucleotide repeats. The AT motif rate was much lower, at 1.1%, while the CG motif was rare at a rate of 0.2%. Similar to the situation of dinucleotide repeats, trinucleotides repeats were not evenly distributed and all possible motifs were present. The most abundant types were AGG (18.9%), AAG (11.8%), AGC (10.5%) and GAT (8.2%), while the other sixteen motifs were present at lower frequencies. For the

tetranucleotides, thirty-one motifs were detected, being AACG and GTTT motifs the most abundant with a rate of 26.1% and 9.2% respectively. Finally, twenty-five different pentanucleotide and nineteen hexanucleotide motifs were found with AATAG, CTATT, CTGGTT and ACATGG motifs present at higher frequencies.

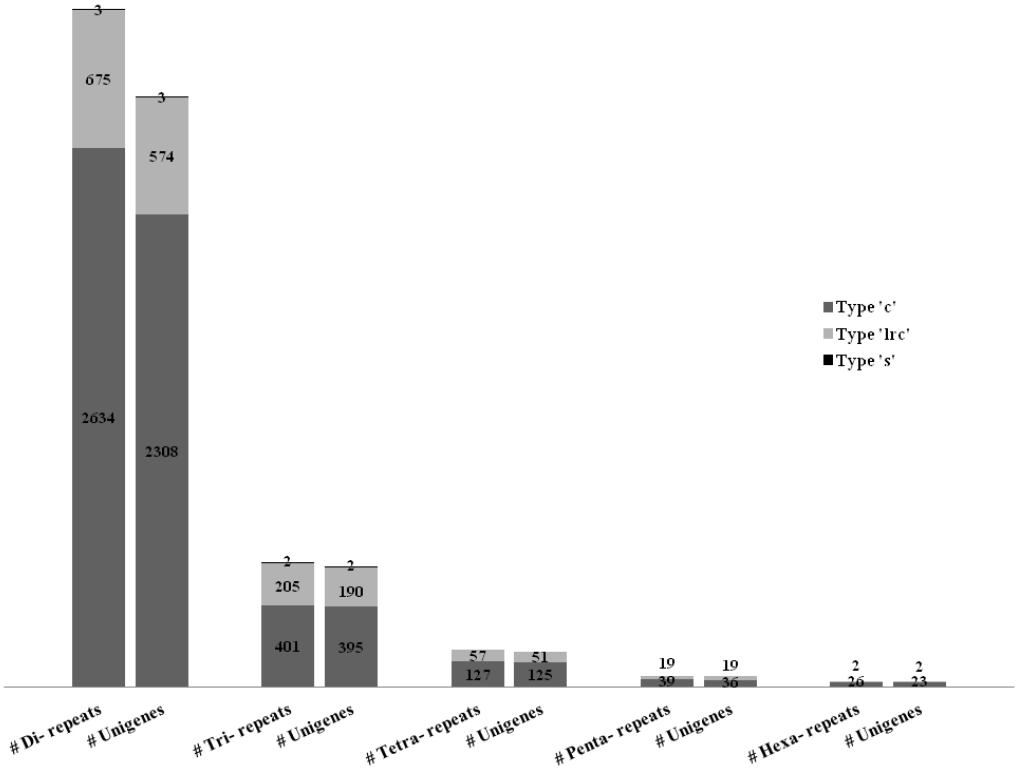


Figure 8 - Frequencies of microsatellites among all unigenes types of *S. pavo*.
Left bars correspond to the total number of microsatellites found for each type in each unigene category and right bars correspond to the total number of non redundant unigenes containing that type of microsatellite. Dinucleotide (Di-), Trinucleotide (Tri-), Tetranucleotide (Tetra-), Pentanucleotide (Penta-), Hexanucleotide (Hexa-), Singletons ('s'), Repetitive contigs ('lrc') and Contigs ('c').

3.2.2 Microsatellite selection

The high proportion of microsatellite containing unigenes within the peacock blenny unigenes offered a unique opportunity for identifying potential type I markers. For this, BLASTX results for the 3,670 unique microsatellite containing unigenes were retrieved from the previous searches and showed that only 779 unigenes were from known genes (Figure 9). To determine the usefulness of the 853 microsatellites harboured by those unigenes, the flanking sequences of the microsatellites were evaluated for primer design. Microsatellites at both ends of the unigenes (incomplete microsatellites) or for which one of the flanking sequences had less than 18 bp (minimum length considered for primer design), were named microsatellites without flanking regions and therefore not useful for marker development.

Table 4 - Characterization of microsatellite motifs in the unigenes of *Salaria pavo*.

	Dinucleotide		**Trinucleotide**		**Tetranucleotide**		**Pentanucleotide**		**Hexanucleotide**	
	Motif	Total Count	Motif	Total Count	Motif	Total Count	Motif	Total Count	Motif	Total Count
	AC	1322	AGG	115	AACG	48	AATAG	10	CTGGTT	6
	GT	1282	AAG	72	GTTT	17	CTATT	10	ACATGG	4
	AG	347	AGC	64	GGAT	12	AAACT	4	CATGGT	2
	CT	320	GAT	50	ATCC	11	AATTC	4	AATACT	1
	AT	36	AAC	46	AAAC	10	AGTTT	4	ACACGG	1
	CG	5	CTT	45	CTGT	10	AATAC	2	ACACTC	1
			CCT	41	CATT	9	ACTAG	2	ACAGCT	1
			GCT	41	GCGT	9	ATCGT	2	ACCACG	1
			GTT	39	CTTT	6	CCCGG	2	ACCAGG	1
			ACC	26	AATG	5	GAATT	2	ACCTGG	1
			ATC	26	ACAG	5	GTTTT	2	ACTGGT	1
			ATT	12	ACGC	5	AAAAT	1	CAGGTT	1
			CCG	8	CAGT	5	AAATT	1	CATCTT	1
			ACG	5	AATC	4	AAGCT	1	CATGCT	1
			GGT	5	ATGT	4	ACAGT	1	CCGGCT	1
			AAT	4	ACAT	3	ACGGT	1	CCTGGT	1
			AGT	4	ATCT	3	ACTAT	1	CTACTT	1
			ACT	2	AAAG	2	ACTCT	1	CTAGTT	1
			CGG	2	CGTT	2	ATCTT	1	GAGAGT	1
			CGT	1	GAGT	2	CCTTT	1		
					GATT	2	CTGGT	1		
					AACT	1	CTTTT	1		
					ACCG	1	GAGGT	1		
					ACCT	1	GATTT	1		
					ACGG	1	GCTCT	1		
					ACTC	1				
					AGAT	1				
					ATTT	1				
					CCGT	1				
					CCTT	1				
					GGTT	1				
Total	3312 (79.0%)		608 (14,5%)		184 (4.4%)		58 (1.4%)		28 (0.7%)	

This analysis only left 637 and 2,569 microsatellites from known and unknown genes respectively for possible marker development (Figure 9).

MSATCOMMANDER in the primer output results gave a list of 1,546 pair of primers, which corresponds to 36.9% of the microsatellites and to 1420 unique unigenes (262 and 1158 with and without BLASTX results respectively). For the remaining microsatellites, primers could not be designed due to short or inappropriate flanking regions for the primer parameters chosen.

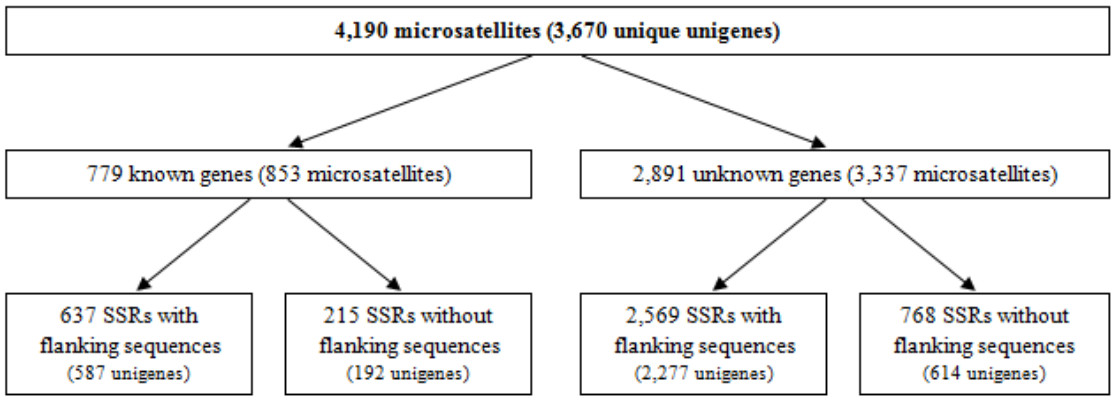


Figure 9 - Estimation of the proportion of type I microsatellites and useful microsatellites among the microsatellites identified from peacock blenny unigenes.

Mining for microsatellites in an EST database offers many possibilities, being one of them knowing *a priori* if a microsatellite will be polymorphic when applied in the population. To achieve this, all of the 4,190 microsatellites were *in silico* mined for polymorphism. In this analysis it wasn't possible to determine the polymorphism *in silico* for 1,428 microsatellites due to being incomplete (28.2%) or only having coverage of one read for the microsatellite region (71.8%). No more than 733 microsatellites (17.5%) were found to be polymorphic *in silico*, of which 727 were dinucleotides and only 6 were trinucleotides.

Without counting with incomplete microsatellites, the average number of repeat units varied between 6.7 repeats for trinucleotides (20.1 bp) and 8.7 repeats for pentanucleotides (43.5 bp, Table 5).

Table 5 - Average repeat length and the most observed length for each microsatellite type.

	Dinucleotide	Trinucleotide	Tetranucleotide	Pentanucleotide	Hexanucleotide
Average	8.3±2.5	6.7±1.1	8.1±3.3	8.7±3.1	7.3±1.5
Mode	6	6	6	7	6
Total micro.	2,990	583	135	54	26

The average read coverage for all microsatellites (without counting with the incomplete ones) was 3.7 ± 9.8 reads with a mode of 2 reads per microsatellite. For dinucleotides and

trinucleotides the read coverage was very disperse, and so the existence of a large number of microsatellites (279 and 63 microsatellites respectively) that could be considered outliers with a read coverage higher than the majority of the other microsatellites (Figure 10). The other microsatellite types hadn't a read coverage so disperse, and that is reflected on only having between 1 to 9 microsatellites with a higher coverage. All microsatellite had as their median 2 reads, except for hexanucleotides that it was 2.5 reads, meaning that half of the microsatellite data considered for this analysis had a read coverage of 1 or 2 reads for the microsatellite region.

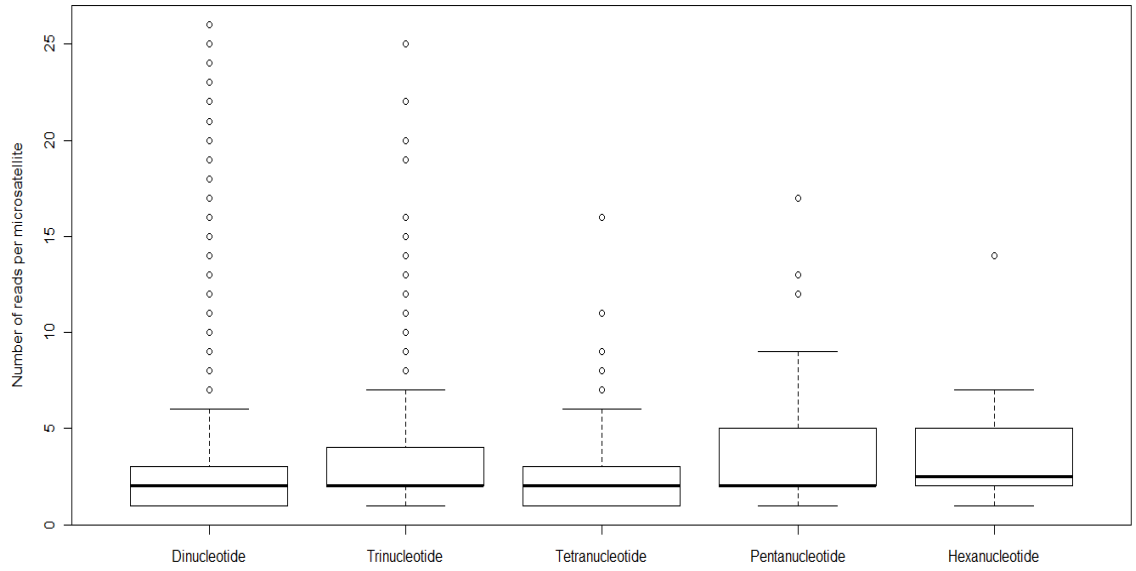


Figure 10 – Box plot distribution of the number of reads per microsatellite type (R statistics).
The maximum number of reads found for dinucleotide and trinucleotide microsatellites is not represented in this graphic due to the discrepancy of the values (349 and 175 reads respectively). For dinucleotide type, 279 microsatellites (not all depicted in this figure) had a higher coverage from the rest of the other 2711 microsatellites. In the case of trinucleotide type, 63 microsatellites (not all depicted in this figure) had a higher coverage than the other 520 microsatellites. For tetranucleotide, pentanucleotide and hexanucleotide types were found 9, 3 and 1 microsatellites respectively with higher coverage than the rest of other 126, 51 and 25 microsatellites respectively. For this analysis were only considered complete microsatellites (with 5' and 3' flanking regions) and reads that covered in totally the microsatellite region on the unigene.

The first strategy used to isolate microsatellites is an attempt to isolate type I markers that have GO terms assigned and are polymorphic *in silico*. For the 733 microsatellites polymorphic *in silico* (17.5%), the maximum number of alleles observed was in two dinucleotide microsatellites with 4 alleles each one (Figure 4, Figure 11). With the remaining criteria of the first strategy, the available number of markers decreased rapidly to only 97 microsatellites with a pair of primers available (Figure 12).

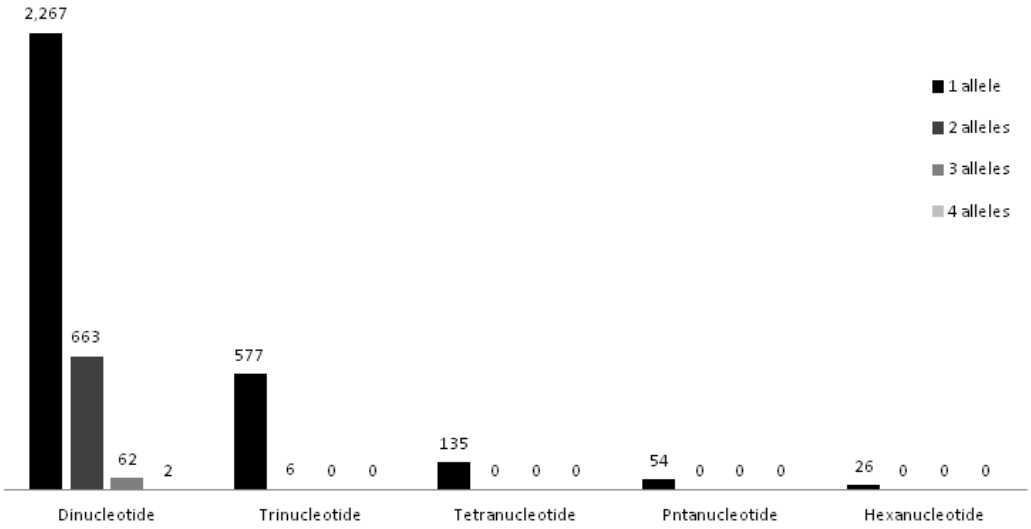


Figure 11 - Distribution of the microsatellites types per number of alleles observed *in silico*.
For the dinucleotide microsatellites was observed up to 4 alleles per loci, for the trinucleotide microsatellites was observed only loci with one or two alleles and for the other types of microsatellites no variation was observed.

From the final set of 97 polymorphic microsatellites, comprised of only dinucleotides (Figure 12), 33 microsatellites were selected based on the BLASTX e-value, the non-redundancy of the unigene and the quality of the flanking regions. These microsatellites had an average read coverage of 21.7 ± 60.6 reads and an average repeat length of 8.5 ± 2.0 units.

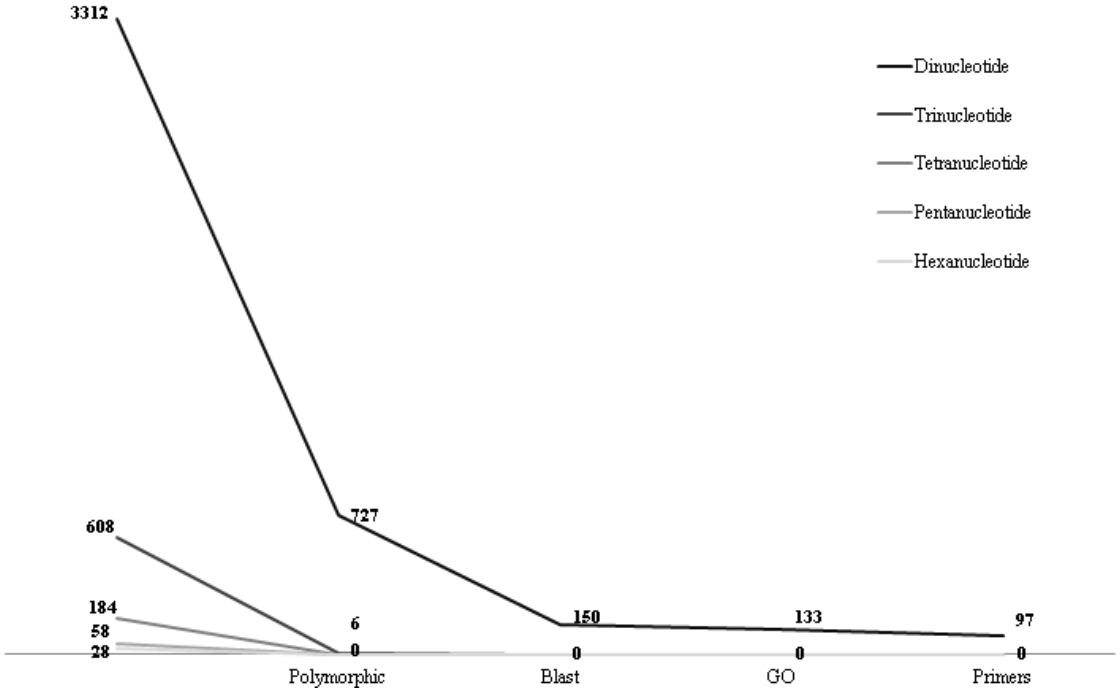


Figure 12 - Results of the microsatellite annotation following the first strategy.
For each microsatellite type, the narrowing of the results for each parameter is depicted.

Following the second strategy, 29 microsatellites were selected based only on the length of the repetition. Of these microsatellites ten were dinucleotides, six were trinucleotides, eleven were tetranucleotides, one was a pentanucleotide and two were hexanucleotides. They had an average read coverage of 3.4 ± 2.8 reads and an average repeat length of 12.0 ± 3.3 units.

3.2.3 Microsatellite application

When the 63 selected primers were PCR checked on one peacock blenny DNA sample, 38.1% (*n*=24) led to different or multiple PCR products and 61.9% (*n*=39) resulted in PCR products of expected size.

Three microsatellites developed in *Lipophrys pholis*, that previously showed to amplify in *Salaria pavo* [55], were also checked by PCR. After PCR conditions were optimized, only one microsatellite gave a clear and strong band when visualized by electrophoresis (6-6). One of the other microsatellites (7-3), although having a unique band but weak, was also used due to his high variability.

Forty-one microsatellites were applied on peacock blenny individuals DNA and sent to fragment sizing. Of these, three microsatellites had mononucleotide variation and six microsatellites had multiple peaks or had not a clear peak in the target region. Other two microsatellites need to be furthered investigated because their fragment variation wasn't in accordance to the corresponding type of microsatellite. One of the microsatellites used from *Lipophrys pholis*, the 7-3 microsatellite, had to be discarded due to its instability in the PCR amplification, perceptible when analysing the results from fragment sizing.

Twenty-eight microsatellites developed from *Salaria pavo* unigenes and one microsatellite developed in *Lipophrys pholis* were successfully characterized in all individuals used from the three locations (Table 6). Of the 28 microsatellites, seventeen were selected using the first strategy and eleven using the second strategy. The microsatellite sequences haven't yet been published, but they are present in Appendix C with the annotation results (when available), coverage depth, number of alleles and the microsatellite location in the gene (when possible). The microsatellite location was obtained by analysing the alignments of the unigene containing the microsatellite aligned to the sequence obtained in the BLASTX results.

For the population of Culatra all but five dinucleotide microsatellite loci were found to be polymorphic (Spavo15-Spavo19; Table 6). The number of alleles ranged from 2 to 12 (average 4.83) alleles per locus and the observed and expected heterozygosities from 0.05 to 0.85 and 0.05 to 0.79, respectively. The average number of alleles per locus and the expected heterozygosity were the highest in microsatellite loci isolated using the second strategy (6.5 and 0.62) than in microsatellites isolated using the first strategy (3.54 and 0.40). In this population, only Spavo14 (p-value <0.001) and Spavo25 (p-value <0.05) loci departed from Hardy-

Weinberg equilibrium expectations most probably because of heterozygote deficit (homozygote excess). The deviation of HW expectation in the first loci is significant and it can be explained by the presence of null alleles or stuttering leading to scoring errors. No other loci were detected with null alleles. Two of the possible pairwise comparisons between loci showed to be significantly in linkage disequilibrium (p-value<0.01: Spavo05-Spavo08 and Spavo08-Spavo25).

For the Formentera population all but eight microsatellite loci were found to be polymorphic (Table 6). The number of alleles ranged from 2 to 5 (average 2.43) and expected heterozygosities from 0.33 to 0.93. No deviation to HW equilibrium and linkage disequilibrium was found.

In the case of the Borovac population all but twelve microsatellite loci were found to be polymorphic (Table 6). The number of alleles ranged from 2 to 5 (average 2.65) and expected heterozygosities from 0.33 to 0.87. No deviation to HW equilibrium and linkage disequilibrium was found.

Microsatellite 6-6 developed in *Lipophrys pholis* was described as having 16 alleles and a perfect microsatellite of 21 repeat units ($(GA)_{21}$). In the case of *Salaria pavo* this perfect microsatellite was broken through substitution of nucleotides and converted to only one polymorphic microsatellite of 11 repeat units (Figure 13).

```
Lipophrys_pholis_micro_6-6_seq      GCAACACTCAGTCAGGCATCAAGGGTTTTAGCATTGATTAGCATTGCTGT
Salaria_pavo_micro_6-6_seq          GCAACACTCAGTCAGGCATCAAGGGTTTTAGCATTGATTAGCATCACTGT
                                    *****************************************  ****

Lipophrys_pholis_micro_6-6_seq      TAGCATGATTCGCATCTCTTGCTTCTGTCTAGTCAGACCTCAAGGGAGTA
Salaria_pavo_micro_6-6_seq          TAGCATGATTCGCATCTCTTGCTTCTGTCTAGTCAGGTCTCGAGGGAGTA
                                    ***********************************  *** ********

Lipophrys_pholis_micro_6-6_seq      AATGTCCCTGGGTGAAGGGCAGAGCACAGAGAGAGCG-AGAGAGAGGGAG
Salaria_pavo_micro_6-6_seq          AATGTCTCTGGGTGAAGGGCAGAGCACAAATAACATGGAGAGAGCGAGAG
                                    ****** ********************* * *    * ****** * ***

Lipophrys_pholis_micro_6-6_seq      AGGGAGAGAG-GGGGGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAGAG
Salaria_pavo_micro_6-6_seq          ATAGAGAGAGCGAGAGAGAGAGAGAGAGAGATAGATAGAGAGAGCGAG
                                    *  ******* * * ****************** *** ******** ***

Lipophrys_pholis_micro_6-6_seq      AGAGAGATGAAAAGGAAGACTGCGAGAGCAAATCCTTGTGGAGAACTGGT
Salaria_pavo_micro_6-6_seq          AGAGACATGAAAAGGATGACTGCGAGAGCGAATCCTTGTCGAAACCTGGT
                                    ***** *********** ************ ********* ** * *****

Lipophrys_pholis_micro_6-6_seq      TGCACACATCCTGTTGCTATAGCAACCTGTGCAGACCAGCTTCTCTACTG
Salaria_pavo_micro_6-6_seq          TGCACACATCCTGTTGCTATAGCAACCTGTGCAGACCAGCTTCTCTACTG
                                    ************************************************

Lipophrys_pholis_micro_6-6_seq      AATGGAGC
Salaria_pavo_micro_6-6_seq          AATGGAGC
                                    ********
```

Figure 13 - Sequence alignment of a cross-species marker (microsatellite 6-6).
Nucleotide sequence alignment between *Lipophrys pholis* and *Salaria pavo* sequences from microsatellite 6-6 [55], performed using ClustalW [71]. Stars under each nucleotide indicate a perfect match and highlighted gray region indicates the microsatellite region.

Table 6 - Primer sequences, characteristics (type and number of repeats), amplification conditions (optimized annealing temperature) and diversity (number of alleles per locus) in the 28 microsatellite loci developed from unigenes in *Salaria pavo* and one microsatellite adapted from *Lipophrys pholis* [55] for individuals from 3 populations.
Loci are sorted from di- to pentanucleotides with descending numbered names.

Locus	Repeat motif	Primer Sequence (5'-3')	T_a (°C)	CULATRA					FORMENTERA					BOROVAC				
				Size (bp)	*n*	*k*	H_O	H_E	Size (bp)	*n*	*k*	H_O	H_E	Size (bp)	*n*	*k*	H_O	H_E
Spavo01[§]	$(GT)_6$	F: FAM-CACCTCGAACAGTTGGCTTC R: GCTGCATTAGCCCAGATCC	58	387−397	20	3	0.30	0.27	387	3	1	-	-	385-387	3	2	0.33	0.33
Spavo02[§]	$(GA)_8C(GA)_4$	F: FAM-CCCTGGCTGATGTGACTCC R: ACTCTCCAGGTGTAAGGCAC	61	250−258	20	5	0.25	0.28	254-256	3	2	0.00	0.53	262-268	3	2	0.67	0.53
Spavo03[§]	$(AC)_6...(GT)_6$	F: FAM-GCACAAGTCGGCACTCAAG R: GCCAAGCCGAGTATGAAGC	60	229−237	20	4	0.50	0.58	235[a]	3	1	-	-	235[a]	3	1	-	-
Spavo04[§]	$(AC)_6$	F: FAM- CCCACGTCTGTTCAGTTGAC R: GGAGTTGGCACATTCCGTG	58	259−266	20	3	0.40	0.45	264	3	1	-	-	258	3	1	-	-
Spavo05[§]	$(AC)_9$	F: FAM-ATCAGCGCGAAACACATCG R: ACTGCACTCAAGTCAAAGCC	56	185−189	20	3	0.55	0.52	183-189	3	3	1.00	0.73	195-199	3	2	0.67	0.53
Spavo06[§]	$(TG)_8$	F: FAM-GCTGGTCGATGGCAGAATG R: GCGTCGGAAATACCGTTCC	58	295−297	20	2	0.05	0.05	295-297	3	2	0.68	0.53	297	3	1	-	-
Spavo07[§]	$(CA)_{11}$	F: FAM-CACGACAGCTGGTCTCAAC R: GGGCTCACCAGTCCCATTC	58	331−337	20	3	0.35	0.42	331-333	3	2	0.33	0.33	331-337	3	3	0.33	0.73
Spavo08[§]	$(CA)_9$	F: FAM-CGTGACTTCATGGCAAGGG R: TGTGTGGAAACGATATGTGC	58	221−235	20	7	0.75	0.79	225-231	3	4	0.67	0.80	231-237	3	2	0.33	0.33
Spavo09[§]	$(AC)_8$	F: FAM-CGCTAAAAGGAGGCAACATC R: ACAGCGACGAGCTTCATCTT	61	196−200	20	3	0.10	0.10	198	3	1	-	-	198-200	3	2	1.00	0.60
Spavo10[§]	$(CA)_9$	F: FAM-AGAGTAGGGGTCCGTCGATT R: TGGCAGTGAGAAAGTGCAAG	61	137−141	20	3	0.10	0.19	153-161	3	5	1.00	0.93	173	3	1	-	-

Table 6 (continued)

Locus	Repeat motif	Primer Sequence (5'-3')	T_a (°C)	CULATRA					FORMENTERA					BOROVAC				
				Size (bp)	*n*	*k*	H_O	H_E	Size (bp)	*n*	*k*	H_O	H_E	Size (bp)	*n*	*k*	H_O	H_E
Spavo11[§]	$(CT)_9$	F: FAM-GGTAGCGAGAGACGCAGAAG	62	232−234	20	2	0.60	0.43	230-232	3	2	0.33	0.33	232	3	1	-	-
		R: GGTAGACCAGCGGTCTGAAG																
Spavo12[§]	$(AC)_7G(AC)_9$	F: FAM-GCTGTAAAACTGCGTGGACA	61	179-203	20	5	0.60	0.56	195-201	3	2	0.33	0.33	203-213	3	4	1.00	0.80
		R: GGACGTGAACCTGGAGAAGA																
Spavo13[§]	$(AC)_{10}$	F: FAM-CCTCGCAGCAGTAACTCAGA	61	136-146[b]	20	3	0.60	0.59	136-144[b]	3	2	0.33	0.33	138[b]	3	1	-	-
		R: TCCGTCTATGGAGGCTAACG																
Spavo14	$(AC)_{17}$	F: GGGGATCGAAATGTTTCACA	59	246-260[**]	20	5	0.40	0.75	248-250	3	2	0.00	0.53	256-260	3	2	0.67	0.53
		R: CCACATGGAACCAACTTCCT																
6-6	$(GA)_{11}$	F: HEX-GCAACACTCAGTCAGGCATC	59	296-312[b]	20	5	0.65	0.64	310-312[b]	3	2	1.00	0.73	290-292[b]	3	2	0.33	0.33
		R: GCTCCATTCAGTAGAGAAGC																
Spavo15[§]	$(AC)_6T(AC)_4$	F: FAM-CATGGCCTATCTGTTCCGC	58	240	20	1	−	−	240	3	1	−	−	240	3	1	-	-
		R: AGACCAACATCCCAGTCGC																
Spavo16[§]	$(AC)_5T(AC)_5$	F: FAM-GTTCAGGATGACCCGGTGG	56	168	20	1	−	−	164-168	3	2	0.33	0.60	164	3	1	-	-
		R: TGTGTATGAGTTCCTGCCC																
Spavo17[§]	$(TC)_7$	F: FAM-TGTCAAGCTCACAGCGAC	56	216[a]	20	1	−	−	216[a]	3	1	−	−	216-218[a]	3	2	0.33	0.33
		R: ATGGCACCCATGCTTCAGG																
Spavo18[§]	$(GA)_7$	F: FAM-CCATGACCAACTACGACGAG	62	175	20	1	−	−	175	3	1	−	−	175	3	1	-	-
		R: GGAGCTTAGGTCGCTCACC																
Spavo19[§]	$(CA)_7$	F: FAM-ACCTTCCAGCCTACGAGAGC	62	170	20	1	−	−	170-172	3	2	0.67	0.53	164	3	1	-	-
		R: TGTGTCAGGAGTAGGCAGACC																
Spavo20	$(AGC)_{10}$	F: FAM-TGCTCGGCTCTACGGTTC	60	209-239	20	8	0.60	0.50	227-233	3	3	0.33	0.73	221-224	3	2	0.33	0.33
		R: CCCTCACAGAGTTCACGGG																

Table 6 (continued)

Locus	Repeat motif	Primer Sequence (5'-3')	T_a (°C)	CULATRA					FORMENTERA					BOROVAC				
				Size (bp)	*n*	*k*	H_O	H_E	Size (bp)	*n*	*k*	H_O	H_E	Size (bp)	*n*	*k*	H_O	H_E
Spavo21	$(AATG)_{14}$	F: FAM-TGTGTTGGTTTGAGACGGC	60	298-330	20	8	0.85	0.79	298-314	3	3	1.00	0.73	302-338	3	4	0.67	0.80
		R: CCTCAAAGACATTGGATGCG																
Spavo22	$(ATCC)_{14}$	F: HEX- GGCAGAAGGAAACCTGGAC	61	139-187	20	9	0.85	0.77	143-189	3	3	0.33	0.73	195-211	3	4	1.00	0.87
		R: GGCCCTTGAAACTCCACTCT																
Spavo23	$(CATT)_8$	F: HEX-CGACCCATTTCGGTTACAAG	61	245-269	20	6	0.75	0.72	249-257	3	3	0.67	0.73	245	3	1	-	-
		R: GAACGAGTAACGTGATGCTGA																
Spavo24	$(CTGT)_9$	F: FAM-GCTCCAACAGAGATAAAACGCTCT	62	170-182	20	4	0.30	0.27	178-182	3	2	0.33	0.60	174-178	3	2	1.00	0.60
		R: TCACTGTAGGAACACGGGAAT																
Spavo25	$(CTGT)_{10}$	F: HEX-GAGTGAGCCGGAGTGTTCTG	62	232-244 *	20	3	0.30	0.55	232-236	3	2	0.33	0.33	228-232	3	2	0.67	0.53
		R: GGCTAAACTGTGGCTGCCTA																
Spavo26	$(GTTT)_8$	F: HEX-CACGTTGCCAATTCCAGTAG	59	212-220	20	3	0.40	0.38	212	3	1	-	-	204	3	1	-	-
		R: GAAGACGACAACCACTCTCAG																
Spavo27	$(AAAC)_{13}$	F: FAM-GAGCTGGCGTTTCCCAAATA	59	169-232	20	12	0.80	0.76	185-189	3	2	0.33	0.33	177-220	3	5	1.00	0.93
		R: ACGGCGTAGTGAGCATGTTG																
Spavo28	$(CTATT)_{10}$	F:HEX-GCAGAGTGACAATAAAGGACGA	59	292-328	20	7	0.75	0.68	302-307[a]	3	2	1.00	0.60	307-333[a]	3	3	0.33	0.73
		R: CCACAAGGCTCAGTTTGACA																

Ta (°C) – annealing temperature; Ho – observed heterozygosity; He – expected heterozygosity; *k* – number of alleles; *n* – number of individuals tested;

"FAM" or "HEX" at the 5' –end of the primer indicate FAM- or HEX-labelled primer.

§ – Strategy 1; a – Mg=1,0mM; b – Mg=1,75mM

Hardy-Weinberg expectation deviations, *p<0.05, **p<0.001

All microsatellite loci were used for reconstruction of genetic distance trees for individuals and populations (Figure 14) in order to see the phylogenetic relations between the samples used in this study. The reconstruction of the genetic distance tree for individuals with D_{AS} distance revealed a clear clustering of all individuals in accordance to their population origin. It was possible to observe two major clades (all descendants with a common ancestor), where the first includes the individuals of Culatra and Formentera populations, and a second clade comprised of only individuals from Borovac population. In the first major clade it was also possible to distinguish two other clades, each one composed of only individuals of Culatra or Formentera populations. The reconstruction of the genetic distance tree for populations with D_C distance held the same results to the tree of individuals. Borovac population presented a more distant position from the other two populations, while Formentera and Culatra populations were most closely related between each other.

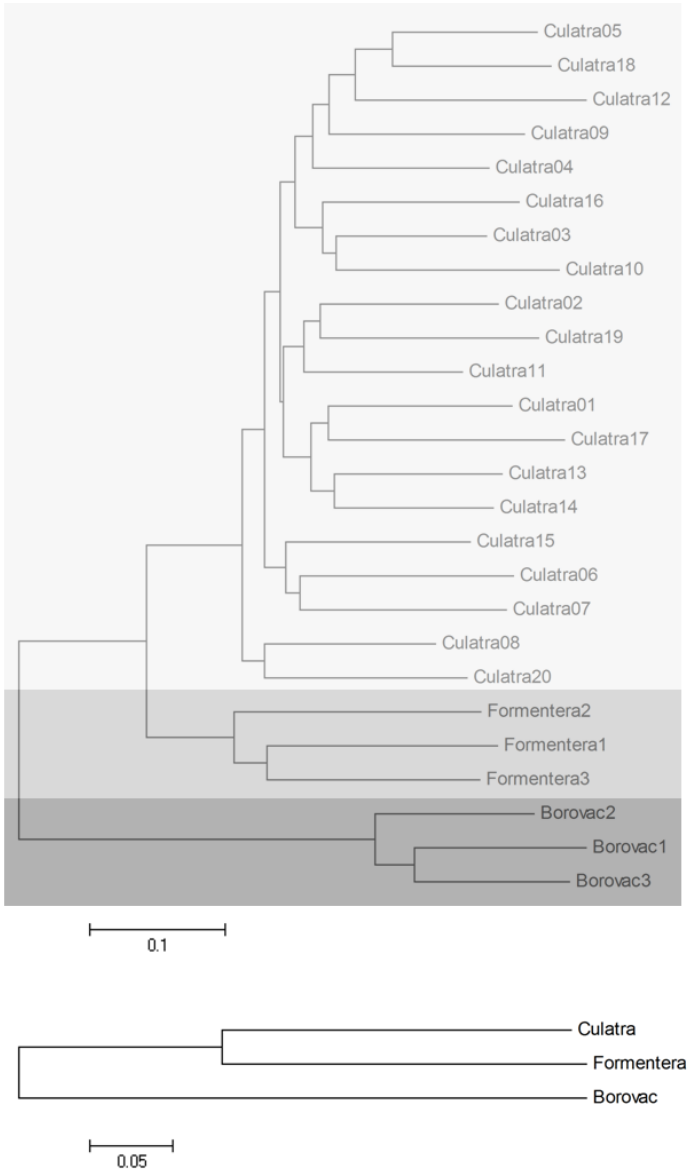


Figure 14 - Trees of individuals calculated with D_{AS} genetic distance (top tree) and of populations according to Dc genetic distance (bottom tree), based on 29 microsatellite loci. Differences in gray colour highlight the individuals according to their population of origin (light gray - Culatra, intermediate gray - Formentera, dark gray - Borovac).

4. Discussion

4.1 Annotation results

A substantial number of ESTs were generated in this project, currently representing the only sequences available for *S. pavo*. Over 640,000 ESTs were sequenced from a normalized library constructed with mRNA pooled from several tissues (see method 2.1.1). The normalization and pooling strategies were used to increase probability of identifying unique and a diverse set of transcripts. Normalization success was confirmed by the fact that most unigenes contained only two reads. Roche's GS FLX technology generates large numbers of ESTs long enough to be informative in the absence of a reference genome [34,35] which allows a *de novo* transcriptome assembly of these sequences into longer contiguous sequences more representative of a complete transcript. The assembly resulted in 62,038 unigenes, a remarkably high number compared to the amount of genes found in the five sequenced fish genomes (20,000-26,000). One major reason for the high number of transcripts is the short length of the reads, which may result in several assembled contigs and singletons for each gene. Another important factor is the process of alternative splicing, where one single gene can result in several transcripts by combining and removing different parts (usually exons). This mechanism is common in higher eukaryotes, including in fish [72], and the number of expressed transcripts is therefore substantially higher than the number of genes in the genome.

The average read length of this assembly (452 bp) is comparable to those from other studies using the same technology and similar approaches (average = 427 bp, range 312-531 [32,33,73,74]). The average number of reads obtained per contig was 9 and it is in the range of previous studies cited (average = 22, range 8-40), and similar to those where the cDNA library was normalized.

Gene ontology is widely used to standardize representation of genes across species and provides a controlled vocabulary of terms for describing gene product [39]. It was possible to match 31% and 26% of the peacock blenny unigenes to the nr protein and GO databases and annotate 21% of the unigenes, a figure comparable to those of other studies using similar approaches (average = 45%, range 18%-72% [32,33,73,74]).

BLASTX top-hit species distribution of gene annotations showed highest homology to *Tetraodon nigroviridis* protein sequences, followed by *Danio rerio* (Table 2), both with complete genomes available. *Tetraodon nigroviridis* is the closest species with better annotation in the public databases to *Salaria pavo*, with a Genome Annotation Score (GAS) of 9080 (*Danio rerio* has only a GAS of 5848). The Integr8 GAS lodged at the European Bioinformatics Institute (EBI), attempts to provide a very rough measure of how well characterized a genome is. The GAS score measures how much annotation has been associated, on average, with the

protein coding genes of a genome, taking into account a variety of criteria. The majority of top BLAST results (80%) were attributed to fish species, meaning that the functional annotation obtained for *Salaria pavo* sequences have biological significance. These results indicate a high level of phylogenetic conservation of the peacock blenny gene content compared to these species.

The annotation results for *Salaria pavo* can be greatly increased in the near future due to the recent finalization of the preliminary gene annotation of *Oreochromis niloticus* (Nile tilapia) genome, a species belonging to the same order.

4.2 Microsatellite mining and annotation

A total of 3,670 peacock blenny unigenes were identified harbouring 4,190 microsatellites (Table 3). The frequency of unigene-microsatellite (EST-SSRs) obtained for *Salaria pavo* of 6.75% is well in the range reported for other fish species, which ranged from 1.5% in *Xiphophorus* to 11.2% in channel catfish *Ictalurus punctatus* ([25,75]). This value can only be considered an empiric value because of the lack of standardized choices of search criteria for detection of microsatellite types among the various projects. The dinucleotide repeat motifs were the most abundant microsatellites in peacock blenny, accounting for 79% of the microsatellites, value similar to the observed in other species (ranged from 47% for *Oryzias latipes* to 78% for *Xiphophorus*), followed by the tri-, tetra-, and penta- and hexanucleotide microsatellites (Figure 8). The distribution of dinucleotide repeats was similar to what has been found in other fish species except for *Fundulus* in which AT was the most abundant motif [75]. The rare CG/GC motif was found in the peacock blenny unigenes. In relation to trinucleotide repeats, all motifs were found in *Salaria pavo* unigenes and the most abundant motif was AGG corresponding to nearly 19%. For the remainder microsatellite types they were present at lower frequencies (0.7 – 4.4%).

Off all the software available for *in silico* mining microsatellites (for a review see [76]), MSATCOMMANDER was chosen due to its simplicity, manipulable output and inbuilt primer design module.

Although all microsatellites developed from ESTs projects represent potential type I markers, only those who have their unigene identified via BLAST searches and its polymorphism validated can be converted to type I marker. In peacock blenny microsatellites, 637 microsatellites from known genes and 2,569 microsatellites from unknown unigenes could be considered for further analysis as potential type I and II respectively (Figure 9).

When considering the number of primers obtained with MSATCOMMANDER, 1546 pairs, and the estimated number of microsatellites with flanking regions for primer design, 3,206 microsatellites, it gives a difference of 1,660 microsatellites without a pair of primers.

This discrepancy can be explained by the poor flanking regions or the high stringency of the default parameters for primer design in MSATCOMMANDER.

In this work one of the things that was pursued was to explore the possibility to *in silico* determine the polymorphism of a microsatellite loci before it was applied in a population to assess its 'real' polymorphism. This was only possible due to how the cDNA library sequenced was created. Pooling mRNA from several individuals (see method 2.1.1) enabled the caption of their genetic variance, in particular their microsatellite alleles, that when seen together in a particular locus allowed to the see the microsatellite polymorphism. This can be extremely helpful in order to decrease the number of monomorphic microsatellites, especially if the species in question has a low genetic variability.

Thirty-four percent of the *Salaria pavo* microsatellites weren't characterized in relation to their polymorphism (*in silico*) due to be incomplete (28.2%) or only having a coverage of one read in the microsatellite region (71.8%). This number is significant, not only for this analysis but also to their usefulness for primer design. Therefore additional data is necessary to make these microsatellites useful for primer design (former) and polymorphism evaluation *in silico* (both). Considering the parameters of the first strategy adopted here, of the initial 733 microsatellites found to be polymorphic *in silico* (17.5%), only 97 microsatellites could be used.

Recently, a study using a similar approach of this work for mining polymorphic microsatellites *in silico* was published [22]. It isn't clear how many sequences they accessed to obtain an *in silico* number of alleles, but their main target were unigenes with one of two GO terms assigned of interest to the study, 'imuno' and 'growth'.

4.3 Microsatellite application

Although of the lower rate of success in microsatellite implementation, only nearly 45%, this value is in the range of applicability of other works developing microsatellites from unigenes (ESTs-SSRs; average = 62%; range 45% - 76% [22,26,77,78]). Even so, it was obtained more polymorphic microsatellites than in Li *et al.* [77] (15/26 microsatellite) and with less null alleles present than in Kim *et al.* [78] (6/17 microsatellite loci). Amplification failure may be due mainly to primers annealing onto neighbouring exonic regions separated by intron(s) or straddling an exon-intron junction preventing them to bind to genomic DNA template. So, in the future, before designing the primers the exons and introns must be investigated in order to avoid some possible failed amplifications. In this study, one of the motives for not taking this approach was because of the lack of a reference genome, so the decisions would have to be based in predictions of the open reading frames (ORFs). It is also worth noting that usually when the primers are applied and genotyped in a set of samples, more

microsatellite primers reveal to fail as it was the case of this work. Forty-one microsatellites were sent to analysis, but only twenty-nine microsatellites were scorable (see results).

Individuals from the populations of Formentera and Borovac were used in this work in order to verify if the microsatellite primers worked in all DNA samples and not only on those of Culatra where they were designed.

All but five microsatellites (Spavo15-Spavo19) showed to be polymorphic in the Culatra population, being all these microsatellites isolated following the first strategy. This may be explained by the small number of individuals used for genotyping so that by increasing this number the polymorphism could be detected. When looking to the results of these microsatellites obtained in the other populations, only two microsatellites (Spavo15 and Spavo18) stay without being confirmed as polymorphic.

Microsatellites have been classified according to the type of repeat sequence as perfect, imperfect, interrupted or composite [79]. In a perfect microsatellite, the repeat sequence is not interrupted by any base not belonging to the motif; in an imperfect one, there is a pair of bases between the repeated motifs that does not match the motif sequence; in an interrupted microsatellite, there is a small sequence within the repetitive sequence that does not match the motif sequence, while in a composite microsatellite, the sequence contains two adjacent distinctive sequence repeats [79]. Although all microsatellite loci were developed as perfect microsatellites, five microsatellites cannot be considered perfect. Spavo02 and Spavo15 already were imperfect in the unigene from where they were isolated. The decision to use them was due to the fact that the second microsatellite had a small repetition length ($\leq$ 4 repeats) non-polymorphic and that wouldn't interfere with the assessment of polymorphism, as it was confirmed latter. Spavo03 has two microsatellites varying in length, but the second microsatellite (GT) wasn't detected when confirming the amplified sequence by the respective pair of primers and so amplified by mistake. It cannot be considered an interrupted microsatellite because the motifs are different, although they have a short sequence of twenty-one nucleotides separating them. Spavo12 and Spavo16 were perfect microsatellites when they were isolated, but once their sequences were confirmed by sequencing the nucleotide introducing the imperfection to the microsatellite was present. In the case of Spavo16 only the microsatellite before the interruption was confirmed to be polymorphic, but in the case of Spavo12 the two microsatellites were polymorphic. Spavo27 and Spavo28 loci showed to have an insertion of a nucleotide on some alleles that was later confirmed by sequencing.

Comparing the results obtained using two different strategies some conclusions can be drawn. The rate of failure was higher for microsatellites isolated using the second strategy (10/29 = 34.5% success) than using the first strategy (17/33 = 51.5% success). On the other hand, the second strategy revealed to be more effective in getting microsatellites with higher rates of polymorphism (average of 6.5 alleles per locus) than the first strategy (average of 3.54

alleles per locus). Petit *et al.* [80] suggested that microsatellite loci with more repeats generally show higher mutation rates (probably because DNA slippage[10] increases in proportion to the number of repeats), which could explain these results.

Merging these two strategies may help improving the polymorphism results and at the same time develop type I markers. This may have the disadvantage of perfect microsatellites become imperfect microsatellites, as happened in Spavo12 and Spavo16. Zhu *et al.* [81] showed that imperfections accumulate in microsatellite sequences and tend to persist. Their accumulation reduces slippage and may lead to the loss of the characteristic microsatellite features of runs of perfect repeats but improve the microsatellite stability [82], important for microsatellites harboured by genes. This accumulation of imperfections is sometimes referred to as 'the death of a microsatellite' [83]. A good example of the accumulation of imperfections is the microsatellite locus 6-6 in *Salaria pavo*. A perfect microsatellite locus of twenty-one repeats described in *Lipophrys pholis* was reduced to a perfect microsatellite of only eleven repeats in *Salaria* (Figure 13).

In relation to a previous mentioned study [22], that used only microsatellites that were polymorphic *in silico*, they obtained a lower rate of polymorphism (61% vs. 82%). Although using a strategy that *a priori* guarantees that microsatellite loci will be polymorphic, it seems to be more productive to select in the first place a microsatellite in relation to its repeat length and then to other parameters of interest, when the rate of polymorphism is of interest. As it has been already stated, the length and location of an SSR motif is an important factor in determining its usefulness as a marker, the longer the repeats, the higher the probability a marker will be polymorphic [80]. As it happened in Hoffman *et al.* [22], some microsatellite when applied revealed not polymorphic, and in these cases it is necessary to increase the number of individuals genotyped or test in other populations of the same species, as it was done in this work.

Genotyping individuals with capillary electrophoresis has proven to be the technique with the lower averaged mean difference (0.73 bp) between the estimation and the actual size of an allele [84], proving to be a good tool in ecological studies. Despite this, it is necessary to have some precautions when analysing data obtained using fluorescence. One problem associated with this approach is the issue of dye-induced mobility shift ('die shift'), due to variability of mobility during electrophoresis between the different fluorescent dyes [58]. Using different dyes for the same locus can introduce large size differences (2.07 bp to 3.68 bp) and affect

[10] DNA polymerase slippage can occur during DNA replication or repair, in which one DNA strand temporarily dissociates from the other and rapidly rebinds in a different position, leading to base-pairing errors and continued lengthening of the new strand and an increase in the number of repeats (i.e. additions) in the allele if the error occurs on the complementary strand or a decreased number of repeats (i.e. deletions) if the error occurs on the parent strand [79].

microsatellite genotyping without a consistent pattern for standard corrections [57,58]. Switching between the same dye was demonstrated to introduce marginal differences in allele scoring (0 to 0.25 bp) [58]. This cannot be an issue for genotyping if one dye is always used for a particular microsatellite locus, but in context of combining data from multiple studies without any correction (when necessary) to the fragment sizes obtained with capillary electrophoresis can lead to differing allele designations. Here, for publication, the allele sizes were corrected to their real size, but for the continuity of this work electrophoretic sizes will be used for the genotyping and statistical analyses. One other problem that can be encountered is 'size shifts'. These occur when the electrophoretic size difference observed between two adjacent alleles (peaks) does not correspond to the exact repeat unit variation of the microsatellite locus [57]. In these cases, to assign the right allele size is necessary to sequence more than one sample from the microsatellite locus in order to understand what is happening in relation to the variation. In this study two microsatellites are in this situation and in need of further analyses to understand if they are a case of size shift, a microsatellite with an indel or more than one locus being amplified by the same primers.

The phylogenetic relationships obtained between all individuals are in accordance with the geographic localization of their population of origin. The Borovac population is the most isolated geographically (Figure 5) and therefore with greater genetic distances and different allele content (Table 6) in respect to Culatra and Formentera populations. Regarding Culatra and Formentera populations they have a lower genetic distance between them, and so they are grouped in the same major clade. These results highlight the usefulness of the selected microsatellites for *S. pavo* population genetic studies.

4.5 Future Directions

Like many organisms that are of great interest in behavioural research, blennies and their close relatives lack complete genome sequences and most other genetic tools and resources. With this work, twenty-eight novel microsatellite loci, of which seventeen can be considered Type I marker, were developed and possibly be used in cross-species amplification in the blenny species.

From this set of microsatellite loci, the most interesting will be selected for paternity studies in *Salaria pavo*, in order to estimate the rate of success of parasitic males in fertilizing nesting males eggs.

As it was mentioned earlier, the two tactics used here complement each other and so it would be interesting to automate the process of *in silico* mining for polymorphism and retrieving of the gene annotations, since it is becoming increasingly common to combine the massive sequencing technologies and microsatellite mining for marker development. The

microsatellites of *Salaria pavo* have been manually curated in this work and so this data can be used to test in a small application to be developed.

5. References

1. Zander CD (1986) Blenniidae. In: Whitehead PJP, Bauchot, M-L, Hureau, J-C, Nielsen, J and Tortonese , E, editor. Fishes of the North-Eastern Atlantic and the Mediterranean: Paris, UNESCO. pp. 1096-1112.

2. Patzner RA, Seiwald M, Adlgasser M, Kaurin G (1986) The reproduction of Blennius pavo (Teleostei, Blenniidae) V. Reproductive behavior in natural environment. Zoologischer Anzeiger 216: 338-350.

3. Almada VC, Goncalves EJ, Santos AJ, Baptista C (1994) Breeding ecology and nest aggregations in a population of Salaria pavo (Pisces: Blenniidae) in an area where nest sites are very scarce. Journal of Fish Biology 45: 819-830.

4. Almada VC, Goncalves EJ, Oliveira RF, Santos AJ (1995) Courting females: ecological constraints affect sex roles in a natural population of the blenniid fish Salaria pavo. Animal Behaviour 49: 1125-1127.

5. Goncalves EJ, Almada VC, Oliveira RF, Santos AJ (1996) Female mimicry as a mating tactic in males of the blenniid fish Salaria pavo. Journal of the Marine Biological Association of the United Kingdom 76: 529-538.

6. Goncalves D, Matos R, Fagundes T, Oliveira R (2005) Bourgeois males of the peacock blenny, Salaria pavo, discriminate female mimics from females? Ethology 111: 559-572.

7. Mackiewicz M, Porter BA, Dakin EE, Avise JC (2005) Cuckoldry rates in the Molly Miller (Scartella cristata; Blenniidae), a hole-nesting marine fish with alternative reproductive tactics. Marine Biology 148: 213-221.

8. Oliveira RF (2005) Hormones, social context and animal communication. In: McGregor PK, editor. Animal Communication Networks: Cambridge University Press, Cambridge. pp. 481-520.

9. Goncalves D, Alpedrinha J, Teles M, Oliveira RF (2007) Endocrine control of sexual behavior in sneaker males of the peacock blenny Salaria pavo: effects of castration, aromatase inhibition, testosterone and estradiol. Hormones and Behavior 51: 534-541.

10. Avise JC, Jones AG, Walker D, DeWoody JA, Collaborators (2002) Genetic mating systems and reproductive natural histories of fishes: lessons for ecology and evolution. Annual Review of Genetics 36: 19-45.

11. Schlotterer C (2000) Evolutionary dynamics of microsatellite DNA. Chromosoma 109: 365-371.

12. Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. Genome Research 10: 967-981.

13. Field D, Wills C (1998) Abundant microsatellite polymorphism in Saccharomyces cerevisiae, and the different distributions of microsatellites in eight prokaryotes and S-cerevisiae, result from strong mutation pressures and a variety of selective forces. Proceedings of the National Academy of Sciences of the United States of America 95: 1647-1652.

14. Li YC, Korol AB, Fahima T, Beiles A, Nevo E (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Molecular Ecology 11: 2453-2465.

15. Metzgar D, Bytof J, Wills C (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Research 10: 72-80.

16. Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. Ecology Letters 9: 615-629.

17. Chistiakov DA, Hellemans B, Volckaert FAM (2006) Microsatellites and their genomic distribution, evolution, function and applications: a review with special reference to fish genetics. Aquaculture 255: 1-29.

18. Zane L, Bargelloni L, Patarnello T (2002) Strategies for microsatellite isolation: a review. Molecular Ecology 11: 1-16.

19. Slate J, Hale MC, Birkhead TR (2007) Simple sequence repeats in zebra finch (Taeniopygia guttata) expressed sequence tags: a new resource for evolutionary genetic studies of passerines. Bmc Genomics 8.

20. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.

21. Abdelkrim J, Robertson BC, Stanton JAL, Gemmell NJ (2009) Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. Biotechniques 46: 185-190.

22. Hoffman JI, Nichols HJ (2011) A novel approach for mining polymorphic microsatellite markers in silico. PLoS One 6: e23283.

23. Csencsics D, Brodbeck S, Holderegger R (2010) Cost-effective, species-specific microsatellite development for the endangered Dwarf Bulrush (Typha minima) using next-generation sequencing technology. Journal of Heredity 101: 789-793.

24. O'Brien SJ (1991) Molecular genome mapping lessons and prospects. Current Opinion In Genetics And Development 1: 105-111.

25. Serapion J, Kucuktas H, Feng JA, Liu ZJ (2004) Bioinformatic mining of type I microsatellites from expressed sequence tags of channel catfish (Ictalurus punctatus). Marine Biotechnology 6: 364-377.

26. Vogiatzi E, Lagnel J, Pakaki V, Louro B, Canario AVM, et al. (2011) In silico mining and characterization of simple sequence repeats from gilthead sea bream (Sparus aurata) expressed sequence tags (EST-SSRs); PCR amplification, polymorphism evaluation and multiplexing and cross-species assays. Marine Genomics 4: 83-91.

27. Leigh F, Lea V, Law J, Wolters P, Powell W, et al. (2003) Assessment of EST- and genomic microsatellite markers for variety discrimination and genetic diversity studies in wheat. Euphytica 133: 359-366.

28. Mardis ER (2008) Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics 9: 387-402.

29. Kircher M, Kelso J (2010) High-throughput DNA sequencing - concepts and limitations. Bioessays 32: 524-536.

30. Morozova O, Hirst M, Marra MA (2009) Applications of new sequencing technologies for transcriptome analysis. Annual Review of Genomics and Human Genetics 10: 135-151.

31. Mitra RD, Shendure J, Olejnik J, Edyta Krzymanska O, Church GM (2003) Fluorescent in situ sequencing on polymerase colonies. Analytical Biochemistry 320: 55-65.

32. Fraser BA, Weadick CJ, Janowitz I, Rodd FH, Hughes KA (2011) Sequencing and characterization of the guppy (Poecilia reticulata) transcriptome. Bmc Genomics 12.

33. Coppe A, Pujolar JM, Maes GE, Larsen PF, Hansen MM, et al. (2010) Sequencing, de novo annotation and analysis of the first Anguilla anguilla transcriptome: EeelBase opens new perspectives for the study of the critically endangered european eel. Bmc Genomics 11.

34. Torres TT, Metta M, Ottenwalder B, Schlotterer C (2008) Gene expression profiling by massively parallel sequencing. Genome Research 18: 172-177.

35. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, et al. (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Molecular Ecology 17: 1636-1647.

36. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215: 403-410.

37. Pevsner J (2009) Bioinformatics and Functional Genomics. New Jersey, Unitade States of America: Wiley-Blackwell.

38. Gish W, States DJ (1993) Identification of protein coding regions by database similarity search. Nature Genetics 3: 266-272.

39. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25-29.

40. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, et al. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Research 36: 3420-3435.

41. Zhu YY, Machleder EM, Chenchik A, Li R, Siebert PD (2001) Reverse transcriptase template switching: A SMART (TM) approach for full-length cDNA library construction. Biotechniques 30: 892-897.

42. Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, et al. (2004) Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acids Research 32.

43. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research 8: 186-194.

44. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 8: 175-185.

45. Chou HH, Holmes MH (2001) DNA sequence quality trimming and vector removal. Bioinformatics 17: 1093-1104.

46. Smit AFA, Hubley R, Green P (1996-2010) RepeatMasker Open-3.0, http://www.repeatmasker.org.

47. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WEG, et al. (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Research 14: 1147-1159.

48. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28: 27-30.

49. Kotera M, Okuno Y, Hattori M, Goto S, Kanehisa M (2004) Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. Journal of the American Chemical Society 126: 16487-16498.

50. Faircloth BC (2008) MSATCOMMANDER: detection of microsatellite repeat arrays and automated, locus-specific primer design. Molecular Ecology Resources 8: 92-94.

51. Santana QC, Coetzee MPA, Steenkamp ET, Mlonyeni OX, Hammond GNA, et al. (2009) Microsatellite discovery by deep sequencing of enriched genomic libraries. Biotechniques 46: 217-223.

52. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, et al. (2010) Tablet-next generation sequence assembly visualization. Bioinformatics 26: 401-402.

53. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods in Molecular Biology 132: 365-386.

54. Shahbazkia HR, Christen R (2009-2011) Primer Validator version 1.0, http://bioinfo.unice.fr/primervalidator.

55. Guillemaud T, Almada F, Santos RS, Cancela ML (2000) Interspecific utility of microsatellites in fish: a case study of (CT)(n) and (GT)(n) markers in the shanny Lipophrys pholis (Pisces : Blenniidae) and their use in other blennioidei. Marine Biotechnology 2: 248-253.

56. Gagneux P, Boesch C, Woodruff DS (1997) Microsatellite scoring errors associated with noninvasive genotyping based on nuclear DNA amplified from shed hair. Molecular Ecology 6: 861-868.

57. Ellis JS, Gilbey J, Armstrong A, Balstad T, Cauwelier E, et al. (2011) Microsatellite standardization and evaluation of genotyping error in a large multi-partner research programme for conservation of Atlantic salmon (Salmo salar L.). Genetica 139: 353-367.

58. Sutton JT, Robertson BC, Jamieson IG (2011) Dye shift: a neglected source of genotyping error in molecular ecology. Molecular Ecology Resources 11: 514-520.

59. Pasqualotto AC, Denning DW, Anderson MJ (2007) A cautionary tale: Lack of consistency in allele sizes between two laboratories for a published multilocus microsatellite typing system. Journal of Clinical Microbiology 45: 522-528.

60. Rousset F (2008) GENEPOP ' 007: a complete re-implementation of the GENEPOP software for Windows and Linux. Molecular Ecology Resources 8: 103-106.

61. Louis EJ, Dempster ER (1987) An exact test for Hardy-Weinberg and multiple alleles. Biometrics 43: 805-811.

62. Guo SW, Thompson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48: 361-372.

63. Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Molecular Ecology Notes 4: 535-538.

64. Languella O (1999-2011) Populations version 1.2.32, http://www.bioinformatics.org/~tryphon/populations/.

65. Jin L, Chakraborty R (1994) Estimation of genetic distance and coefficient of gene diversity from single-probe multilocus DNA fingerprinting data. Molecular Biology and Evolution 11: 120-127.

66. Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis and models and estimation procedures. American Journal of Human Genetics 19: 233-257.

67. Goldstein DB, Linares AR, Cavallisforza LL, Feldman MW (1995) An evaluation of genetic distance for use with microsatellite loci. Genetics 139: 463-471.

68. Takezaki N, Nei M (1996) Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics 144: 389-399.

69. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4: 406-425.

70. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24: 1596-1599.

71. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and clustal X version 2.0. Bioinformatics 23: 2947-2948.

72. Yeo G, Hoon S, Venkatesh B, Burge CB (2004) Variation in sequence and organization of splicing regulatory elements in vertebrate genes. Proceedings of the National Academy of Sciences of the United States of America 101: 15700-15705.

73. Roeding F, Borner J, Kube M, Klages S, Reinhardt R, et al. (2009) A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (Pandinus imperator). Molecular Phylogenetics and Evolution 53: 826-834.

74. Kristiansson E, Asker N, Forlin L, Larsson DGJ (2009) Characterization of the Zoarces viviparus liver transcriptome using massively parallel pyrosequencing. Bmc Genomics 10.

75. Ju Z, Wells MC, Martinez A, Hazlewood L, Walter RB (2005) An in silico mining for simple sequence repeats from expressed sequence tags of zebrafish, medaka, Fundulus, and Xiphophorus. In Silico Biology 5: 439-463.

76. Sharma PC, Grover A, Kahl G (2007) Mining microsatellites in eukaryotic genomes. Trends in Biotechnology 25: 490-498.

77. Li Q, Liu SK, Kong LF (2009) Microsatellites within genes and ESTs of the Pacific oyster Crassostrea gigas and their transferability in five other Crassostrea species. Electronic Journal of Biotechnology 12.

78. Kim KS, Ratcliffe ST, French BW, Liu L, Sappington TW (2008) Utility of EST-Derived SSRs as population genetics markers in a beetle. Journal of Heredity 99: 112-124.

79. Oliveira EJ, Padua JG, Zucchi MI, Vencovsky R, Vieira MLC (2006) Origin, evolution and genome distribution of microsatellites. Genetics and Molecular Biology 29: 294-307.

80. Petit RJ, Deguilloux MF, Chat J, Grivet D, Garnier-Gere P, et al. (2005) Standardizing for microsatellite length in comparisons of genetic diversity. Molecular Ecology 14: 885-890.

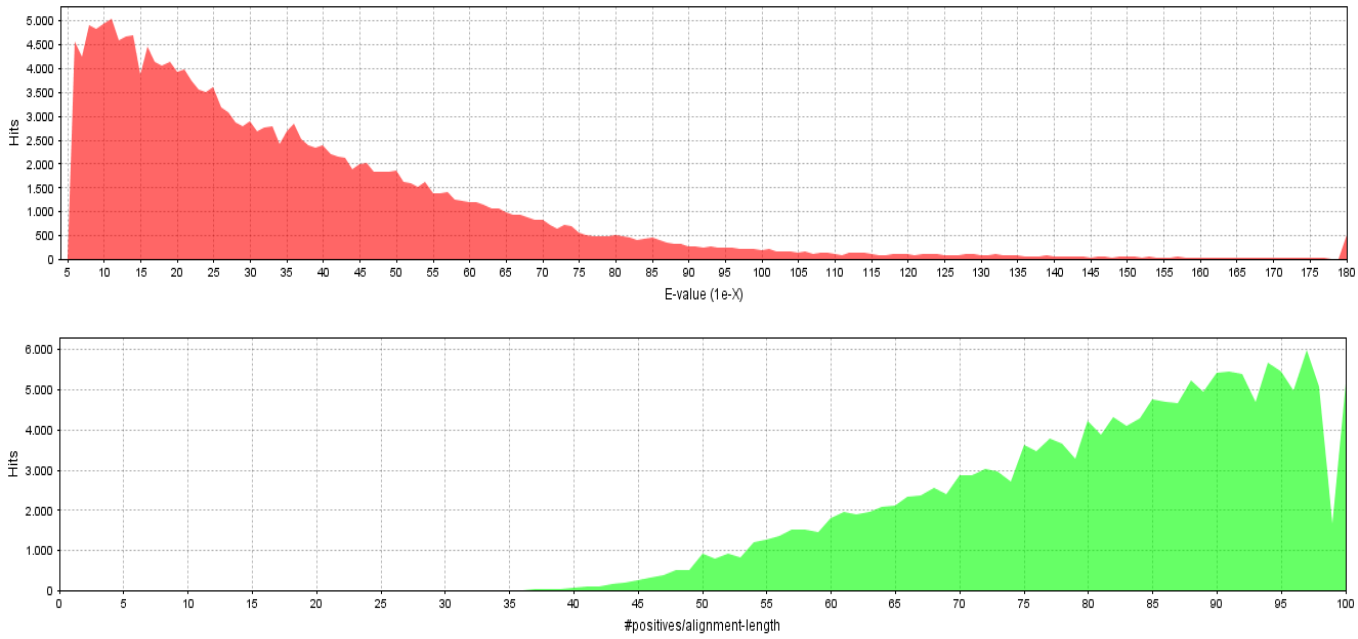
81. Zhu Y, Strassmann JE, Queller DC (2000) Insertions, substitutions, and the origin of microsatellites. Genetical Research 76: 227-236.

82. Bhargava A, Fuentes FF (2010) Mutational Dynamics of Microsatellites. Molecular Biotechnology 44: 250-266.

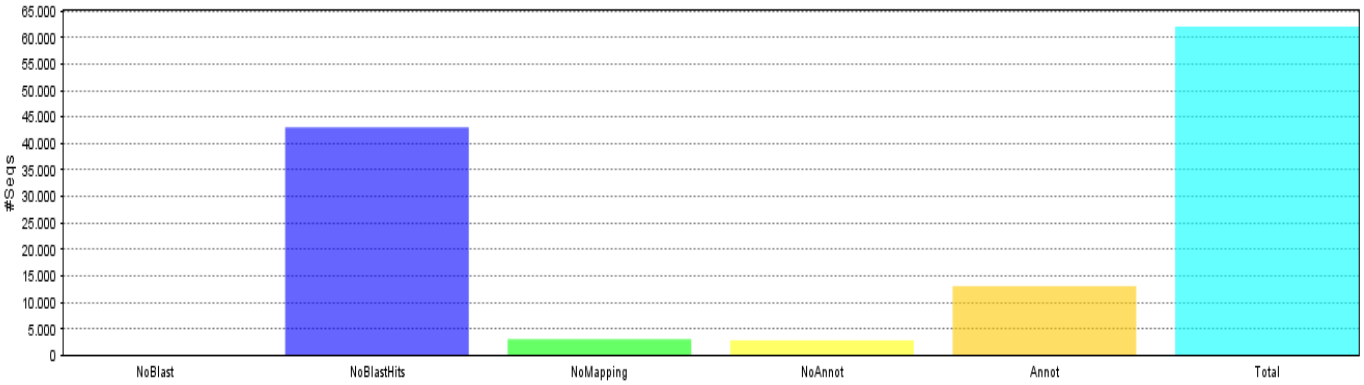
83. Taylor JS, Durkin JMH, Breden F (1999) The death of a microsatellite: a phylogenetic perspective on microsatellite interruptions. Molecular Biology and Evolution 16: 567-572.

84. Vemireddy LR, Archak S, Nagaraju J (2007) Capillary electrophoresis is essential for Microsatellite marker based detection and quantification of adulteration of basmati rice (Oryza sativa). Journal of Agricultural and Food Chemistry 55: 8112-8117.

6. Appendix



Appendix A - E-value (top graphic) and similarity distribution (bottom graphic) for 165,615 hits obtained in BLASTX searches.



Appendix B – Summary of the annotation process of *Salaria pavo* unigenes.

Appendix C - Microsatellite sequences for each microsatellite locus and the respective *in silico* mining, BLAST and gene ontology results.

Locus	Amplicon	Depth of coverage	*In silico* number of alleles	Protein name	e-Value	GO	Repeat location
Spavo01	CACCTCGAACAGTTGGCTTCAGTCATTTTGTTCACYGAACG TATTTTACTTTTATTAAAGCGTCAGAAACAAAAAGTTGTTA GAGGCTGTTGGGGACATTTTTCTGACAAAACCTTCTGCTCC ATGTAAAAMCTYTTTTTCTTTTGAAATCTTTCTGTGAGCGA CGCTYGTTTRATATTTTCCTTTYTTTGTTTACGGCTTGAATT TTGTCACAAACACCGGAAGAGAGTTTTGAATTTAATGTTTC ATGCTTTAATTTAAGGTTTCAGTGGAAACAGTGAAATCAA TATTGATTCTGGAAGAAATAATAATTTTATTTTGCTTTTAG AAAGAGTGTGTGTGTGTTTTCATCTGCTCTCATTAGAAATG AGGATCTGGGCTAATGCAGC	2	2	TAP-binding protein	1,7e-14	**P:** antigen processing and presentation of endogenous peptide antigen via MHC class I **C:** membrane	3'UTR
Spavo02	CCCTGGCTGATGTGACTCCGTCCTCAGGGGTTCCCAACGTG CACCACTTGGCTGGACGGTCACCACGGAGACGACCAGCCA TCACTGCATCTAAAGATCCCAGAGATTCAAACTCCAAAAC ATCTCTTCGGTGAACGAGAGAGTGAGACTGGGCGCAAGAC AATGAGAGAGAGAGAGAGAGCGAGAGAGAACCAGCGCGA ATGTGCGTGTGAGATGATGTGAGCGATGAGAAGTGCCTTA CACCTGGAGAGT	4	2	Ras association domain-containing protein 1	1,5e-21	**P:** intracellular signalling pathway; signal transduction **F:** metal ion binding; receptor activity; protein binding	3'UTR
Spavo03	GCACAAGTCGGCACTCAAGTGACCTTAAACACATCGGCTC TCGGCTCTCTGGGAGGGGTCGACAATAAGCTGATGTGCTT CACCAAACTCTGAAGTCATGGCATACTGTATCGTATCAATC ATAGCAGCTGAACCCGTGCGATACGTTCACGCCATTAGCT AAATTAACACACACACACCTCCTATTTCCATTGTGGCCCGT GTGTGTGTGTGAGTGCTTCATACTCGGCTTGGC	98	2	Hook homolog 1	4,8e-41	**P:** endosome to lysosome transport; spermatid development; early endosome to late endosome transport; endosome organization; protein transport; lysosome organization **F:** microtubule binding; identical protein binding; actin binding **C:** FHF complex; HOPS complex; microtubule	3'UTR
Spavo04	CCCACGTCTGTTCAGTTGACTCAAGTTGATGTGCTGGGTG WASGGCCAGGAGCAACACACACACACAGAAATGTAAATA TCATGTGAAATCGTTTTAATGTCTTTTTTTCTTCTTTTTACT CTATTTAACAGTATTCAGAACAAGGAACGATTTATCTATCT	7	2	HAUS augmin-like complex subunit 5	8,3e-11	**P:** cell cycle; mitosis; spindle assembly; centrosome organization; cell division **C:** cytoplasm; microtubule organizing center;	3'UTR

	GTCGATCTTCTGTACTGCTCTATCATAATATGCATGTCTTT AGATTGTGGGAAGGCACCTGAATATCGAGATGAAAGTGAG CACACGGAATGTGCCAACTCC					spindle; cytoskeleton; microtubule; HAUS complex	
Spavo05	ATCAGCGCGAAACACATCGCCAATGCCGTCCGTCAGACCT TCGCCAACTAAGAACCGCACCTGATTCCCCCCTCCCCCACA CACACACACACACACCATTTCCATCACAAACCAGCGACCA GGCAGGATGACATGAAGACTGGGTAATGTCTCTAGTGCCA CCTTGTGGCTTTGACTTGAGTGCAGT	11	3	Proteasome (prosome, macropain) activator subunit 4	4e-14	**P:** multicellular organismal development; cell differentiation; spermatogenesis **F:** binding **C:** proteasome complex; nucleus	3'UTR
Spavo06	GCTGGTCGATGGCAGAATGTCTGCGTAGGAGAGTCGCTCG TTCACTCGCTCAGCCAAGACTTGTCAACAGCCTACCAAAG ACAGCACTCAGGCTCTTGTTCAATCTAAACCTAGAGGAGG CTCCTGCACTGTTAGCTGCATCTCTTCTCACCGTATGTCTGT GTGTGTGTGTGTGCGCGCGTGTGTTCCCTCGTCCGTCCCTC TCGTTGTTGCCAACCGGTGCTGCTGAGCAGACTTGAAGTCT TTAATTCAGAGCGTTTGTTAAACCTCTTAGCAGGAACGGTA TTTCCGACGC	6	2	Mitogen-activated protein kinase 8 interacting protein 3-like	1,6e-19	**F:** kinase activity	3'UTR
Spavo07	CACGACAGCTGGTCTCAACATGTGTGATTCTGTGTTTCAGT GTGTGAGTCTGGATATGTGCGCCCCGCTATGCGTACATAAT TTTACATGTGGGGTTTTATGTGTAACTTCAGGCAATTCTTC TGTAAACAATTGTGTCAGGGTCCGTTCAAATGACAGAAAC CTGTTTCTTGTTGTCAGACGGCCTGAGATGCCTCAGTCAAA TACACACACGCACACACACACACACACACACGCACTCCCA GCACACCTTCGCAGTTAATCAGTGAAGGAGGAAAACATCT GTCACGGTTGAAGGCCGGAAAATAAACCGAATGGGACTG GTGAGCCC	6	2	Type IV collagen alpha 1 chain	1,1e-23	**P:** cell adhesion **F:** platelet-derived growth factor binding; extracellular matrix structural constituent **C:** collagen type VI; sarcolemma	3'UTR
Spavo08	CGTGACTTCATGGCAAGGGAGCAGAATGAAGTTCGCTTCT CATCCGTGGCGCTTTGTCGCGCTTAAAAGACCAGTCGACG CCAAATCAAACAAATTGTAAACCAAAATGTTCTGGAACAC TCGCATGACAAACACAATCAACAGGCTCAGAGGCTGACCT GTGCACACACACACACACACATCTGCAGACACTATGCACT ATAAAGCACATATCGTTTCCACACA	5	3	Ubiquitin C-terminal hydrolase L1 EC:3.1.2.15, EC:3.4.19.0, EC:3.4.22.0	3,9e-44	**P:**ubiquitin-dependent protein catabolic process; multicellular organismal process; protein deubiquitination; behavior **F:** ubiquitin thiolesterase activity; omega peptidase activity; ubiquitin binding; cysteine-type endopeptidase activity	3'UTR

						C:cytoplasm	
Spavo09	CGCTAAAAGGAGGCAACATCAGTTAAACACACACACACAC ACCCTCTTATTGCAGCTACCCAACTGCTCCTCCTGGCCACT TCAAGAATACTCTCATCAACAAATCATCATGGCAGCCGTG TCAGCTCTCATCATCATRTATCTCGAAGAGTGACGGGGAG AGTGATGCTACAAGTCAAAGATGAAGCTCGTCGCTGT	2	2	Tyrosyl-tRNA synthetase, cytoplasmic EC:6.1.1.1	2,3e-31	**P:** tyrosyl-tRNA aminoacylation **F:** tyrosine-tRNA ligase activity; protein binding; ATP binding; tRNA binding **C:** cytoplasm; nucleus	3'UTR
Spavo10	AGAGTAGGGGTCCGTCGATTTGATCGCCTGCTGCTTCTTCA CGTCTTCAGGTTCAGGTTGCCGATCTTTAACACAGACACGA AACGTTTTATTAGGTCACACACACACACACACACTGACTT GCACTTTCTCACTGCCA	2	2	Sulphite oxidase EC:1.8.3.1	1,1e-13	**P:** response to nutrient; oxidation reduction **F:** electron carrier activity; sulfite oxidase activity; molybdenum ion binding; heme binding; molybdopterin cofactor binding **C:** mitochondrial intermembrane space; cytosol	3'UTR
Spavo11	GGTAGCGAGAGACGCAGAAGGGAGTCGGTCACCAGCTAA CTGGGAAACCACCTCATCCGAAACACCTGCAATCACAAAC TGCTCCCCGGGCTGGTTGGAGACTCTCCAGCCGGACCTTCA GCTCTCTCTCTCTCTCTCTGGGCTGGTTGGCGGACCATGTC GGTCCCGGCGGTGGTCACTGTGCGGCTCATCCGCTCCTTCG AGCACCGCAACTTCAGACCGCTGGTCTACC	14	2	UPF0538 protein C2orf76 homolog isoform 2	2,4e-20	**P:** biologiacal_process **F:** molecular_function	5'UTR
Spavo12	GCTGTAAAACTGCGTGGACAACAACAACAATAAAACTGCG GGACGGCTTTAAAAAGCAGCGACACACACACACACGCAC ACACACACACACACACTCACTCACCGCCGAGAGTCGAGTC TTTCACCTGCAGAGATTTCTTTCCCTGAACTTCTTGAACTA TGGGTGTCATCATCTCTCACCTCTTCTCCAGGTTCACGTCC	6	2	ADP-ribosylation factor	3,2e-29	**P:** vesicle-mediated transport; positive regulation of growth rate; embryonic development ending in birth or egg hatching; small GTPase mediated signal transduction; hermaphrodite genitalia development; molting cycle, collagen and cuticulin-based cuticle; body morphogenesis; nematode larval development; inductive cell migration; intracellular protein transport **F:** GTP binding **C:** Golgi apparatus	5'UTR
Spavo13	CCTCGCAGCAGTAACTCAGACGCATTGTACTGAACACACA CACACACACACACAGACGTCCGCCCGCAGGACTCCGACTA	3	3	Proteasome	4,3e-123	**P:** negative regulation of ubiquitin-protein	3'UTR

	GACATCCAACGTCTTTTAATACAATCAACTCAGCAACGTTA GCCTCCATAGACGGA			activator complex subunit 3 EC:1.4.3.4, EC:1.4.3.6		ligase activity during mitotic cell cycle; regulation of apoptosis; oxidation reduction; anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process; cell adhesion; positive regulation of ubiquitin-protein ligase activity during mitotic cell cycle; amine metabolic process **F:** peptidase activity; quinine binding; copper ion binding; p53 binding; MDM2 binding; identical protein binding; amine oxidase activity; proteasome activator activity **C:** integral to membrane; proteasome activator complex; cell surface; cytoplasm; nucleus; plasma membrane	
Spavo14	GGGGATCGAAATGTTTCACACACCTTTACATTTATTTGTAG CTTTGTCCAGGATGGCTTTTGTTGAAACAATTTTGCCATAC CTAAAATAAGAGGATCAGAAAGACGAGACAATGTYAGTT CACTTCATACCAGACGTRGCCAACTTACAAAGACACACAC ACACACACACACACACACACACACACTGTTGAGCAATATT CATGACATCATTGACAAAAAATGAGAAACAGGAAGTTGGT TCCATGTGG	2	2	-	-	-	-
6_6	GCAACACTCAGTCAGGCATCAAGGGTTTTAGCATTGATTA GCATCACTGTTAGCATGATTCGCATCTCTTGCTTCTGTCTA GTCAGGTCTCGAGGGAGTAAATGTCTCTGGGTGAAGGGCA GAGCACAAATAACATGGAGAGAGCGAGAGATAGAGAGAG CGAGAGAGAGAGAGAGAGAGATAGATAGAGAGAGCG AGAGAGACATGAAAAGGATGACTGCGAGAGCGAATCCTT GTCGAAACCTGGTTGCACACATCCTGTTGCTATAGCAACCT GTGCAGACCAGCTTCTCTACTGAATGGAGC	-	-	-	-	-	-

Spavo15	CATGGCCTATCTGTTCCGCAACATTGCTGTGGATGACTGAG GCAACAAACACACACACACATACACACACTTTAATTCCAG CAAACATTTCCAAGCACACATGCCGAGGACCAAGAATCAA GTTTATTCTAGTTTGTGCTTTGAGCCTCAGTTCCTATTCGTT TTGACCTTATGGACCTGTTGGTAGTACCTCTCTGTGTTCTCT GTGAGTTTTCTATGGAGCGACTGGGATGTTGGTCT	6	2	Phosphoribosyl pyrophosphate synthetase-associated protein 1 EC:2.7.6.1	9,7e-43	**P:** nucleotide biosynthetic process **F:** ribose phosphate diphosphokinase activity; magnesium ion binding	3'UTR
Spavo16	GTTCAGGATGACCCGGTGGATGAGGAAGGTCATTGAGAAG ACAGGAGGAAGTGACGACTGAGCGACAACACACACACTA CACACACACCATATTCACTGGATCCCATGGCACAATGTCG ACACAACGCAATACCACCACCCACCTGCAGGGGCAGGAAC TCATACACA	10	2	Coagulin factor II (Prothrombin) EC:3.4.21.0	1,2e-158	**P:** cell surface receptor linked signaling pathway; platelet activation; positive regulation of protein amino acid phosphorylation; positive regulation of blood coagulation; proteolysis; positive regulation of release of sequestered calcium ion into cytosol; fibrinolysis; positive regulation of collagen biosynthetic process **F:** thrombospondin receptor activity; receptor binding; serine-type endopeptidase activity; calcium ion binding **C:** extracellular space	3'UTR
Spavo17	TGTCAAGCTCACAGCGACGCCCCGTTTTGTTTTCTTTAGGT GTCTCTCTCTCTCTCATTCCTCTTATCGCTCCGCTTCTCAGT GAGGTTTGGGAACATGGCCGTCGACATAGAAATTCCTTGT AACTTTTGGCAAAGTGAGTTCAATGCCATCCAGCTCTGGA GTGAGAACAATCTGACATCCCAGTCGGGAGTTTTCCTGAA GCATGGGTGCCAT	23	2	Adrenodoxin-like mitochondrial precursor	9,1e-36	**P:** transport; electron transport chain **F:** 2 iron, 2 sulfur cluster binding; metal ion binding; electron carrier activity **C:** mitochondrial matrix	3'UTR
Spavo18	CCATGACCAACTACGACGAGGCCGCCATGGCTATCGCCAG CCTAAACGGCTACCGCCTGGGTGACCGCGTGCTGCAGGTT TCCTTCAAGACCAGCAAGCAGCACAAGGCCTGAGAGAGA GAGAGAGGCAGCTGGTACCAGCTCCTTCGRCCTGCGGGGT GAGCGACCTAAGCTCC	3	2	ELAV-like protein 3 (Hu-antigen C)	1,2e-153	**P:** cell differentiation; nervous system development **F:** RNA binding; nucleotide binding **C:** ribonucleoprotein complex	3'UTR
Spavo19	ACCTTCCAGCCTACGAGAGCTATGAGAAGCTGAGACACAT GCTGCTGTTGGCCATTCAGGAATGCTCGGAGGGATTCGGA CTGGCTTAAACCACAAGACTCTTTAACGCACACATACACA	3	2	E3 ubiquitin-protein ligase HUWE1	2,7e-113	**F:** acid-amino acid ligase activity **P:** protein modification process	3' UTR

	CACACACACAGACATGCATACACTTATACTAGGTCTGCCT ACTCCTGACACA			EC:6.3.2.0	**C:** intracellular		
Spavo20	TGCTCGGCTCTACGGTTCGGGTGTGCTGACCCGCCGGGGG GCCGGAGTGCAGAAGTCTAAAGCAGCTGAAGCCAAGCCA GCAGCAGCAGCAGCAGCAGCAGCAGCTGAGAGGAAG CACTCC	15	1	-	-	-	-
	CGTTGGTCTGCCCAGCAAACTAGCAGGCTCCAGAAGAAGC AGGCCGAGGCAGTGCTCCCTCCTTTCTCCTGGAAGTTTCCA GAAGACCCCGTGAACTCTGTGAGGG						
Spavo21	TGTGTTGGTTTGAGACGGCAGCAGCTGATTGAGGAACAGA AGGTTTTAGATWTTTAAATGAGGAAAGACGTGAGCAGAC AGGAAACAGGAAGGTGACGTCTGAAAACACAAAATACTT CGGTATGGTCGTAAAGTTTGAGAGCGAAATGAACCACAGT CAAAACAGATCAGTGTAGAAAGGTTAGGAAAAGAGAAAT TAACATTAGTGTCTAAAAAAGATAGATATGGGAAAGGAAT GAATGAATGAATGAATGAATGAATGAATGAATGAATGAAT GAATGAATGAATGAATGTGAGGAGTCGATCGCATCCAATG TCTTTGAGG	2	1	-	-	-	-
Spavo22	GGCAGAAGGAAACCTGGACAGGTCACCAGTCAATTAGTGG GCAAACACCGGCCATCCATCCATCCATCCATCCATCCATCC ATCCATCCATCCATCCATCCATCCATCTCTTCTTTTTC TGTAACTCCTTTATCAAGAGTGGAGTTTCAAGGGCC	5	1	-	-	-	-
Spavo23	CGACCCATTTCGGTTACAAGAATCACATGTGTGAGAACTT GTACCATGAACTGAATGCTCAAAAATTCATCACCTGCACT GATACTGAAACTGTGACCGAAATAAACCGCTCATTCATTC ATTCATTCATTCATTCATTCAAGTATACTTGTTGTTGG TGTGTAGTTTCAATCCTATTTGACATCATTACAATATTAAC AGTAATATAAACAGAAACGTGTCAGCATCACGTTACTCGT TC	3	1	-	-	-	-
Spavo24	GCTCCAACAGAGATAAAACGCTCTAGTCTGTCTGTCTGTCT GTCTGTCTGTCTGTCTGTCTGTGGGCTTCCCCGCAGCAGGC ACAACACAGCGCTGTGTACACATATTTCCAGCACTATGAA AGACAGATTTGCGAGAATTGATGAAAATATGAGTAATTCC CGTGTTCCTACAGTGA	2	1	-	-	-	-

Spavo25	GAGTGAGCCGGAGTGTTCTGAGAAAAACTTTGCAGGCAGG GGAGAAAACACACAAACTCCAAGTACAAGACAACTGTAA CAACTGTATGTTTGAGATTTCCTCTAAATGATTGTATTTTC CAGGGTTCAAAGATTCTTTATTTATTCATCTGTCTGTCTGTC TGTCTGTCTGTCTGTCTGTCTGTCCGTCCTTTATTGAATTGG CCCATCAGTAGGCAGCCACAGTTTAGCC	4	1	-	-	-	-
Spavo26	CACGTTGCCAATTCCAGTAGTTAAATATTAGGATATGCGTT AAGTTGGACTCACATTTCTTTACTGTAAATGTTGGGTGCCT AAAAAAAGAACTGATCTGCACACTTGTAAAATGTTYCTTT TGTTTGTTTGTTTGTTTGTTTGTTTGTTTGTTTGTCATACAG CCTTCAAATGTAGGCTTTATTCTAGTTTGTACTGAGAGTGG TTGTCGTCTTC	2	1	-	-	-	-
Spavo27	GAGCTGGCGTTTCCCAAATATTCAACCTACAGCGATGGCT GACATGTCTCTATGTTCACCTTTAACTGTGACAACATTTCT GTTCAAGTTATGTTTTTGCCAAACAAACAAACAAACAAAC AAACAAACAAACAAACAAACAAACAAACAAACATCACCA CTAGTGGCCCAACATGCTCACTACGCCGT	3	1	-	-	-	-
Spavo28	GCAGAGTGACAATAAAGGACGATTCTATTCTATTCTATTCT ATTCTATTCTATTCTATTCTATTCTATTTCAGACAAC GTTGGTGTAGTTGTATCATTGGCATTTCATTAGGAGACTGA TCCTGGATGGCCTGGAGGCAAGGACCCAAGACTTTAAACT TTTCCAAATTTAAGAAAATGTAAAACTCAACCCCCAGAGA GCCAAAGTGTGCTCTCTCGCACCACTTACAACCACTAAAA TGGCCGTCACGGTTCCCTTGTTGCATGCGTCTGTGTCAAAC TGAGCCTTGTGG	2	1	-	-	-	-